UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
Electronic and Computer Engineering
Cycle XXXVIII

TITLE OF THE Ph.D. THESIS

Unmasking Deepfakes, Adversarial Attacks, and Mobile Users in The Deep Learning Era

Scientific Disciplinary Sector

ING-INF/05

Ph.D. Student:	Gianpaolo Perelli
Supervisor	Gian Luca Marcialis
Co-Supervisor	Giovanni Puglisi

Final exam. Academic Year 2024/2025
Thesis defence session: February 2026

The integration of artificial intelligence systems into society is continually increasing, thanks to their rapid solutions that are often more efficient than those of humans. Since these technologies process personal data, they need to be robust and secure. This thesis examines these crucial factors, investigating the vulnerabilities that these systems may encounter in real-world scenarios.

A threat that has inspired significant research activity in recent years is the phenomenon of deepfakes. Their widespread dissemination, ease of generation, and high quality make them among the most dangerous threats. Discrediting a person, spreading false news, and jeopardizing digital identities have become alarmingly accessible. Although studies demonstrate solutions with satisfactory results, their weakness when compression is used cannot be overlooked. Files shared via apps or social media are compressed and resized to facilitate rapid delivery, inadvertently triggering a sort of attack on deepfake detectors. In addition to accidental data sharing issues, it is also important to be aware of those generated by targeted attacks. Focusing on voice disorder detection systems, which are becoming increasingly important in public health, machine learning is increasingly present and crucial in distinguishing between healthy and pathological voices. A significant concern, given the critical nature of personal health information, is that these systems can be deceived by various types of attacks, such as adversarial, evasion, and pitch-based methods. However, attacks are not only about changing a classification; they can also be implemented to obtain confidential information. For this reason, the rapid expansion of smartphones, which collect extensive personal and behavioral data, poses new risks to digital identity. Applications that utilize mobile sensor data for activity, health, or security monitoring may inadvertently enable the identification of individuals or devices.

This thesis tackles these issues by (1) presenting innovative methods to mitigate performance losses in deepfake detectors under compression, (2) providing an in-depth analysis of vulnerabilities in voice disorder detection systems to enhance defenses against attacks, and (3) highlighting potential privacy concerns in human activity recognition systems.

Contents

Chapter 1

Introduction

In recent years, artificial intelligence (AI) and, in particular, machine learning (ML) have seen growth, establishing themselves as fundamental tools across a wide range of applications. Although the origins of ML techniques date back several decades, it was not until 2012, with AlexNet's victory in the ImageNet challenge, that there was exponential growth in the use of neural networks (NN) for AI. Since then, deep learning has experienced rapid growth, and research has increasingly focused on how these techniques could be applied across multiple domains. From biometric recognition systems to medical diagnosis applications, up to the generation of malicious synthetic content known as *deepfake*.

It was only in 2017 that, on the popular forum *Reddit*, an explicit deepfake video was published for the first time, attracting widespread attention. The term deepfake, in fact, originates from the nickname of the user responsible for the publication. This name simultaneously highlights the techniques used to create these videos: a particular type of deep neural network (DNN) based on *Autoencoder* (AE). Although that forum was later banned, the idea, and particularly the underlying technology, had already attracted significant attention: the risks of identity fraud or reputational damage had been clearly exposed. This phenomenon has been rapidly expanding: these are not merely image manipulations produced with professional software, which is often complex to use and time-consuming. Nowadays, mobile applications can easily and quickly generate such fakes. Several DNNs have followed the AEs, such as generative adversarial networks (GAN) and more recent diffusion models (DM).

Thanks to these applications, if in 2022 someone impersonated the mayor of Kyiv using *deepfake* techniques to make video calls with three European capital mayors while promoting pro-Russia arguments[1], in the more recent 2024 case the economic damage for a company was substantial, up to 25.6 billion dollars were

[1] https://www.cnbc.com/2022/06/25/european-mayors-duped-into-calls-with-fake-kyiv-mayor.html

transferred following what appeared to be an explicit request from the chief financial officer during a video call[2].

These episodes demonstrate that this technology has reached a level of quality capable of deceiving even the most discerning eye [93]. Although this technology can be usefully employed in various fields such as entertainment and content productions, it can be considered a serious threat to the reliability of visual content. Therefore, studies focused on the detection of such manipulations are particularly important to prevent fraud, safeguard digital identity and personal reputation, as well as to curb the spread of misinformation. Moreover, the distribution of such content poses an additional challenge: social media platforms and messaging applications compress multimedia files to enable rapid sharing. This compression process inevitably removes information that is essential for the reliable operation of deepfake detectors. In fact, the loss of such information leads to a dramatic drop in performance [149]. Studies in this field are therefore mainly divided into two directions: generalization, i.e., models capable of recognizing types of deepfakes not seen during training; and robustness, i.e., models able to maintain high performance even in the presence of compression [174, 143].

While sharing media involves an automatic process that seriously hinders deepfake detectors, there are other hidden pitfalls in AI models: adversarial attacks. These attacks involve deliberately altering input data to deceive the model into erroneous predictions [182]. These alterations are often imperceptible to people, but they can significantly impact what the model classifies. The risks posed by these attacks on AI models are especially concerning for systems that affect human health, including those relying on audio-based classification [19]. Malicious actors could fabricate audio recordings that falsely indicate medical conditions, potentially resulting in serious consequences, such as job loss or increased insurance premiums, based on fake evidence. Public figures may also be exposed to reputational damage. As with the deepfakes discussed above, fake recordings could be weaponized in cyberbullying or harassment campaigns, distorting public perception. Technologies [194] now enable the convincing manipulation of both visual and audio content, fueling the spread of misinformation and causing harm before experts have the chance to intervene. A growing body of research has explored the use of simple vocal tasks for diagnosing voice disorders. These disorders are defined as alterations in voice quality [13], and in some cases, may be linked to the presence of a tumor. Therefore, developing intelligent systems that enable rapid recognition could save valuable time, especially considering the waiting times for medical appointments. In fact, the integration of machine learning models to support doctors is rapidly growing. Thanks to these systems, real-time diagnoses are

[2]https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnke

possible. In the case of vocal disorders, this approach enables the identification of the presence or absence of the pathology through acoustic analysis of the patient [66].

Among AI systems designed to support human health, Human Activity Recognition (HAR) technologies stand out for their ability to provide tailored suggestions for personal well-being. These systems harness sensor data from mobile and wearable devices to interpret physical movements and have become increasingly prevalent in domains such as healthcare, security, and behavioral analytics [141]. By drawing on information collected from embedded accelerometers, magnetometers, and gyroscopes, HAR solutions can reliably deduce user activities [39], powering applications that span from fitness monitoring to anomaly detection [122, 54]. Yet, the detailed motion data processed by HAR systems raises significant privacy issues, especially concerning the possibility of user or device re-identification [74]. Research has demonstrated that HAR features may unintentionally capture distinctive movement patterns, potentially enabling attackers to track individuals over time [37].

This thesis represents the outcome of a coordinated research effort aimed at developing original methodologies and advancing tools and analyses for more robust and secure AI systems. It begins with an in-depth exploration of Deepfakes, providing an overview of the generation techniques and their evolution in the first part of Chapter 2. It follows an in-depth exploration of deepfake detection, reviewing the state of the art and presenting novel published solutions that enhance detection robustness under real-world conditions. Chapter 3 shifts the focus to machine learning vulnerabilities in voice disorder detection systems, analyzing recent studies and introducing new experimental contributions on adversarial attacks in this domain. Finally, Chapter 4 investigates the privacy risks associated with Human Activity Recognition (HAR) systems, summarizes the most relevant literature, and presents a new analysis of user re-identification using sensor data.

1.1 Contribution

Active contributions were made to research on methods aimed at enhancing the overall robustness of AI systems. This is fundamental for ensuring the reliability and security of multimedia content and AI-driven systems.

Described in Section 2.4.1, a contribution focuses on the recognition of deepfakes applied to images. This work, a natural continuation of my master's thesis, was presented at ICIP 2022 with the title *"Tensor-Based Deepfake Detection In Scaled And Compressed Images"* [31]. We address the deepfake detection problem by exploiting the traces that such manipulations leave on images at the frequency level.

This aspect is crucial, as it allows the extraction of highly discriminative features when the discrete cosine transform is applied to the image. In this study, we introduce for the first time in the state-of-the-art a new type of input that includes the frequency representation of images at different compression scales. We also propose a restyled version of Xception [26], significantly reduced in terms of depth and number of parameters, in order to make it compatible with the new type of input. This approach enables a substantial reduction of the input size compared to the original image, thus allowing the use of lighter NNs.

Another study, adopting a different methodology and currently under review, is presented in Section 2.4.2. In this case, the experimentation was carried out on videos, although the approach is also transferable to images, since it relies on the analysis of individual frames. In this study, we present the use of a Multi-Head module that enables the DNN to learn both artifacts introduced by compression and those generated by manipulations, leveraging the knowledge of the compression level. Furthermore, we introduce two different types of augmentation, *frame-centric* and *video-centric*, addressing an issue that has been largely overlooked in the state-of-the-art, which typically limits itself to training and testing models on only two compression levels [149].

Participation occurred in the *"Face Deepfake Detection Challenge"* in 2022, which resulted in the publication of the article *The Face Deepfake Detection Challenge* in the *Journal of Imaging* [65]. Additionally, involvement in the *"SDFR: Synthetic Data for Face Recognition Competition"* took place in 2024, in collaboration with the BiDA-Lab research group from Universidad Autónoma de Madrid. This collaboration led to a publication in the *IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)* [153].

In Section 3.4.1, the analysis of vulnerabilities in voice disorder detection systems is explored in depth. It presents a series of *white-box* and *black-box* attacks that employ both adversarial and signal-manipulation techniques (such as tone and pitch modifications) to expose and highlight critical weaknesses in these systems. This study was published in *IEEE International Workshop on Information Forensics and Security* (WIFS 2024) under the title *"Vulnerabilities in Machine Learning-Based Voice Disorder Detection Systems"* [131].

Security issues in Human Activity Recognition (HAR) systems are also discussed in Section 4.3. This research led to the publication of the paper *"Unintended Identification in HAR Systems: Evaluating Privacy Risks"* at the *IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering* (MetroXRAINE 2025). The study raises the issue of user re-identification in HAR systems by exploiting HAR-derived features from a single sensor (the accelerometer), even when the performed activity is unknown.

1.1.1 Contextualization within the Scientific Landscape

The contributions on deepfake detection presented in this thesis primarily address the vulnerability of current detectors to video and image compression. In particular, the proposed tensor-based representation and compression-aware architectures introduce solutions that remain effective under realistic compression scenarios, where many state-of-the-art systems exhibit significant performance degradation.

The study on adversarial attacks in voice disorder detection highlights an underexplored security risk in medical AI: the possibility that malicious audio manipulations could lead to incorrect diagnoses. By systematically evaluating both white-box and black-box attack scenarios, this work contributes to the emerging field of adversarial robustness in healthcare-oriented machine learning systems.

Finally, research on HAR investigates the unintended leakage of identity-related information from sensor-based features, exposing privacy risks in these systems. The findings demonstrate that user re-identification may occur even when activity recognition is the primary objective, supporting the need for privacy-aware and privacy-preserving HAR designs.

1.1.2 Statement of Personal Contributions

Building on the overview of the main research directions, this section systematizes the published works, contextualizes their scientific relevance, and maps them to the corresponding chapters of the thesis. Moreover, my individual roles in each publication are detailed to clarify the alignment between the thesis content and my personal research. Table 1.1 summarizes the publications in chronological order, detailing for each work the main personal contributions, the related thesis sections, and the corresponding research area.

Table 1.1: *Mapping between publications, personal contributions, related thesis sections, and research area.*

Publication	**Main Personal Contributions**	**Related Sections**	**Research Area**
Tensor-Based Deepfake Detection in Scaled and Compressed Images (ICIP 2022) [31]	Development of tensor-based input representation, redesign of the Xception architecture, experimental implementation, analysis of compression effects, manuscript writing.	Section 2.4.1	Deepfake
The Face Deepfake Detection Challenge (Journal of Imaging 2022) [65]	Application and adaptation of the proposed tensor-based model within the challenge framework, experimental implementation and evaluation.	Discussed only for context in Chapter 2	Deepfake
SDFR: Synthetic Data for Face Recognition Competition (FG 2024) [153]	Evaluation and testing procedures.	Briefly discussed in Section 2.2.1	Synthetic data
Vulnerabilities in Machine Learning-Based Voice Disorder Detection Systems (WIFS 2024) [131]	Implementation of adversarial and non-adversarial attacks, development of white-box and black-box strategies, evaluation of model vulnerabilities, interpretation of risk implications.	Section 3.4.1	Adversarial attacks
Unintended Identification in HAR Systems: Evaluating Privacy Risks (MetroXRAINE 2025)	Development of an experimental protocol for user re-identification, feature analysis from accelerometer data, privacy risk quantification.	Section 4.3	HAR privacy
Robust Deepfake Detection in Compressed Videos with Scalable Network Strategies (under review)	Architectural design and implementation of Multi-Head and Multi-Branch architecture, development of frame- and video-centric augmentations, experimental evaluation, result analysis, manuscript writing.	Section 2.4.2	Deepfake

Chapter 2

Deepfakes

This chapter, entirely dedicated to the phenomenon of deepfakes, aims to provide a general overview of the topic. It begins by presenting the technologies underlying their generation, tracing their evolution from the first AE-based approaches to the most recent models. It then lists the various categories of deepfakes, which are distinguished by the target of the attack and the manipulation techniques employed. The discussion continues with a review of studies focused on combating deepfakes, including an in-depth analysis of state-of-the-art deepfake detectors. The chapter concludes by presenting our original studies and techniques, which have contributed to enriching the literature in this field.

2.1 Introduction to Deepfakes and Their Technologies

Since the birth of digital photography, techniques have been used to enhance image quality through the use of software. As software has become increasingly complex, it has become possible to manipulate image characteristics in ways imperceptible to the human eye. Professionals in the field still use these techniques. Generally, the common operations are: *splicing*, when one crops something from an image and pastes it into another; *inpainting*, when one wants to remove something from an image; *copy-move* is the classic duplication of something in the same image. In these cases, the identification of a fake occurs by detecting errors in shadows, illumination, or perspective. Technological advancements and continuous research have contributed to the exponential development of deep learning models, yielding remarkable achievements in various domains, with the generation of synthetic content being a notable example [178]. As reported in recent studies [93], humans are unable to reliably distinguish between face images generated by state-of-the-art methods and those that are genuine. These methods utilize DNNs, which have experienced exponential growth in various areas since 2012: AEs, GANs, and DMs have proven particularly important for generating multimedia content. These new technologies enable rapid image and video manipulation without requiring extensive specialization in the field, with incredible results. After a brief introduction to how the technologies work [60], we'll look at the different types of deepfakes present in the literature.

2.1.1 Autoencoders (AE)

Autoencoders are DNNs designed to compress and then reconstruct input data [69]. To do this, they pass the input through a structure called a *bottleneck*. This structure forces the model to extract only the essential information. In this process, autoencoders identify hidden patterns in the data and compress large amounts of information into compact representations. The model learns by minimizing the difference between the input and its reconstruction. This minimization aims to reduce the reconstruction error, which serves as the typical loss function. All autoencoders follow a fundamental structure with three main parts:

- *Encoder*: This part compresses the data. As an input passes through the encoder's layers, its dimensionality is reduced step by step, keeping the most relevant information.
- *Bottleneck*: This is the most compact part of the architecture. The generated code here represents a reduced and abstracted version of the original input. The code must hold enough key details to allow the decoder to reconstruct the entire story.
- *Decoder*: The decoder performs the opposite operation of the encoder. It takes the compressed code and *decompresses* it to reconstruct the original data.

The earliest deepfakes were made using a shared autoencoder (AE) with two separate decoders, one for each person [170]. During training, thousands of images of Person A and Person B were used. The encoder learned to map any face into a latent representation. Decoder A was designed to reconstruct a face as Person A, regardless of input, while Decoder B was designed to reconstruct a face as Person B. At inference time, to create a deepfake, frames were extracted from a video of Person A, and each face was encoded into the latent space. Instead of passing the latent code through Decoder A (which would yield a reconstruction of Person A), the latent code was fed into Decoder B. Decoder B then generated an output that preserved the pose, expression, and alignment of Person A, but with the facial identity features of Person B.

2.1.2 Generative Adversarial Networks (GAN)

While AEs have been adapted for generating artificial content, GANs were specifically designed for this purpose: to generate new, original, and realistic data that closely resembles the dataset used in training. A GAN learns the intricate details of human features from a dataset of real images. In contrast, an AE primarily reconstructs existing faces and lacks the ability to innovate beyond the given examples. This highlights the significant leap that GANs represent. Their potential lies not just in mimicry but in genuine originality. Introduced in 2014 with the paper [61], these DNNs overcome several limitations of AEs. GANs consist of two networks: a Generator, G, and a Discriminator, D. They both engage in an adversarial game where the generator tries to deceive the discriminator by generating data similar to that in the training set. The Discriminator, on the other hand, strives not to be deceived and learns to distinguish fake data from real data. This dynamic interaction allows GANs to learn and work with complex data, such as audio, video, or image files. The generator produces images from random noise and gradually refines its ability to generate realistic images. Random noise, the input, is sampled from a uniform or normal distribution and then enters the generator to produce an image. The generated images, together with real images from the training set, are fed into the discriminator, which learns to differentiate between the two. The output D is the probability that the input is real: if the input is real, D equals 1; if it is generated, D approaches 0. The power of this technology has enabled the creation of entirely synthetic faces and serves as the foundation for several models [139], which, when combined with autoencoders (AE), facilitate the generation of different types of deepfakes that will be discussed in the following paragraphs.

2.1.3 Diffusion Models (DM)

DMs are today among the most advanced architectures for image, audio, and video generation, often surpassing GANs in both quality and stability. Introduced in 2020 [70], they rely on a different approach compared to GANs: instead of generating an image from random noise in a single step, DMs learn to progressively remove noise from a corrupted input, step by step, until a new image emerges. The training process comprises two main phases: the *Forward* and the *Reverse* processes. In the Forward phase, controlled noise is gradually added to a clean image, degrading it until it becomes pure noise. In the Reverse phase, a convolutional neural network (CNN), typically a U-Net [146], an architecture originally developed for image segmentation, is trained to predict and remove the noise added at each step. Through this iterative denoising strategy, DMs can generate highly detailed and coherent content, setting a new standard for synthetic data generation.

Hybrid architectures have emerged as a compelling approach for deepfake generation with DMs. Typically, these systems combine the robust generative abilities of DMs with an encoder network that isolates facial features (such as identity and pose). The pivotal advance, however, lies in integrating natural language as a conditioning tool. In contrast to GANs, where control is achieved through the intricate manipulation of latent vectors, recent efforts combine DMs with vision-language models like CLIP [147], thereby moving control to a semantic level. This enables users to intuitively guide generation or editing through direct text prompts (*e.g.*, "make this person smile") [86, 142].

2.2 Common Deepfake Categories

The state-of-the-art presents several models for generating deepfakes. These models have different objectives:

- Entire face synthesis: fully synthetic faces, that do not exist;
- Identity swap: replacing the faces of some individuals with others;
- Face morphing: blending biometrics of multiple faces;
- Face editing: modifying attributes (*e.g.*, eye/hair color, beard, bangs);
- Expression swap: transferring expressions;
- Lip-sync: transferring speech from text or audio.

2.2.1 Entire Face Synthesis

This technique can generate entirely synthetic and highly realistic faces that do not exist in reality. While it is often useful in the video game industry, it also poses a serious risk, as it can be exploited to create fake social media profiles and spread disinformation.

Generally, this type of *deepfake* is generated using GANs [62]. One of the earliest studies based on this approach is ProGAN [78]. In this case, the authors aimed to improve performance by modifying the synthesis process to progressively increase image resolution during training. The dataset used was CelebA [107], and the implementation is available on GitHub[1]. An alternative approach, StyleGAN [81], employs a version of the *Generator* based on style transfer [72]. For its training datasets, StyleGAN uses the newly introduced Flickr-Faces-HQ (FFHQ)

[1] https://github.com/tkarras/progressive_growing_of_gans

Dataset Name	# Real images	# Fake images	Source
100K-Generated-Images (2019)[81]	-	100K	StyleGAN
10K-Faces (2019) (link)	-	10K	-
DFFD (2019)[32]	-	100K 200,000	StyleGAN ProGAN
iFakeFaceDB (2019)[125]	-	250K	StyleGAN
100K-Generated-Images (2020)[82]	-	100K	StyleGAN2
100K-Generated-Images (2020)[79]	-	100K	StyleGAN2-ADA
GANDiffFace (2023)[116]	-	500K	StyleGAN3 + DM
DCFace (2023)[87]	-	1,2M	DcFace
CelebA	202,599	-	-
FFHQ	70K	-	-
CASIA-WebFace[187]	500K	-	-
VGGFace2[17]	3.31M	-	-
Mega-Face2[123]	4.7M	-	-

Table 2.1: *Entire face synthesis and real images Dataset.*

and CelebA-HQ [78]. The implementation code is publicly available on GitHub[2]. Building upon the original StyleGAN, this model was later improved with StyleGAN2 [82], and subsequently with StyleGAN2-ADA [79]. The latter is particularly interesting since it achieves excellent results even with only a few thousand training images (GitHub[3]). Most recently, the latest version of this approach is StyleGAN3 [80], which is also available on GitHub[4].

The datasets above are primarily used for training NNs to generate synthetic faces. However, of particular interest to the research community is the analysis of generated images to identify common features that can help distinguish them from real ones. Several studies have shown that it is possible to recognize a kind of fingerprint (similar to PRNU in camera images). Nonetheless, this fingerprint tends to vary across different instances of the same GAN. This has motivated the creation of multiple datasets containing synthetic images, as reported in Table 2.1. The same table also lists datasets containing real images, which are essential for comparative studies between authentic and fake samples.

[2]https://github.com/NVlabs/stylegan
[3]https://github.com/NVlabs/stylegan2-ada-pytorch.
[4]https://github.com/NVlabs/stylegan3

In addition to the aforementioned models, recent literature has explored new approaches for generating synthetic datasets that better capture intra-class diversity and inter-class separability using DMs. For example, *VariFace* [186] proposes a two-stage diffusion-based pipeline to synthesize balanced and diverse face recognition datasets, achieving comparable performance to real datasets across various benchmarks. Another recent proposal, *HyperFace* [154], formulates dataset generation as a *packing* problem in the embedding space and synthesizes images conditioned on optimized embeddings, reaching state-of-the-art performance when training face recognition models solely on synthetic data.

Another interesting application of these datasets is to generate recognition systems while avoiding the use of data belonging to real people, which is becoming increasingly difficult to collect due to stringent privacy regulations and copyright issues. A particularly relevant initiative in this direction is the *"SDFR: Synthetic Data for Face Recognition Competition"* [153], organized in conjunction with the *18th IEEE International Conference on Automatic Face and Gesture Recognition* (2024). The competition aimed to evaluate the feasibility of training face recognition models using only synthetic data, thereby directly addressing the privacy, bias, and generalization issues typically associated with web-crawled datasets. It consisted of two distinct tasks: in the first one, the backbone network was fixed and the dataset size was limited, whereas in the second task, participants had full freedom in terms of model architecture, dataset composition, and training pipeline. The submitted systems were evaluated across seven benchmark datasets, including bias assessment across demographic groups. The SDFR competition highlights the growing potential and current limitations of synthetic datasets in replacing or complementing real data for face recognition research and applications. Other state-of-the-art initiatives in this direction include the works presented in [118, 36], which introduce FRCSyn-onGoing, an open competition and benchmark for testing face recognition systems. This initiative specifically focuses on the integration of synthetic data to address challenges related to privacy, bias, and robustness under difficult conditions. Datasets generated with DM-based frameworks (*e.g.* Dual Condition Face Generator, DCFace[5] [87]) or hybrid GAN and DM models (*e.g.* GANDiffFace[6] [116]) are better suited for these competitions. These approaches address previous limitations by producing diverse images of the same identity, similar to real datasets used in face recognition training.

[5] https://github.com/mk-minchul/dcface.
[6] https://github.com/PietroMelzi/GANDiffFace.

2.2.2 Identity swap

This manipulation technique aims to replace the face of a subject in an image with that of another person. Two terms are typically used to describe the original image and the new face: *source* and *target*.

Identity swapping can be an attractive technique in filmmaking, but malicious use may lead to the production of fake pornographic videos, reputational damage, disinformation, financial fraud, and other harmful actions.

Not all identity swap methods rely on DNNs. For instance, the FaceSwap application[7] leverages *computer vision* techniques, while others use *deep learning*, such as the ZAO mobile app[8], another FaceSwap implementation[9], FaceSwap-GAN[10], DeepFaceLab[11], DeepFaceLive[12], FakeApp[13], and DFaker[14].

The swap process typically begins with face detection in an image, followed by the extraction of facial *landmarks*. These landmarks are crucial for aligning the new face with the *target*. The aligned and cropped faces [83] are then fed into an autoencoder to synchronize expressions. Finally, the synthesized face is blended back into the *source* image with refined colors and boundaries to ensure a natural result. In videos, this process is repeated frame by frame.

For this type of manipulation, several state-of-the-art datasets contain both real and manipulated videos. Table 2.2 shows the first group of datasets (up to FaceForensics++ (FF++) [149]) struggles to provide a sufficient level of generalization, mainly due to the limited variation in poses, lighting, and contrast [174, 173, 68]. These early fake videos, generated with primitive deepfake methods, often contain visible artifacts. By contrast, more recent datasets offer a significantly higher degree of realism.

2.2.3 Face Morphing

Facial *morphing* aims to extract biometric features from two or more faces to generate a new synthetic identity. The resulting sample closely resembles the original inputs, potentially compromising facial recognition security systems [152, 177]. This technique is generally applied to frontal images and involves determining correspondences between faces using *landmark points*. The faces are then warped

[7]https://github.com/MarekKowalski/FaceSwap.
[8]https://apps.apple.com/cn/app/id1465199127.
[9]https://github.com/deepfakes/faceswap.
[10]https://github.com/shaoanlu/faceswap-GAN.
[11]https://github.com/iperov/DeepFaceLab.
[12]https://github.com/iperov/DeepFaceLive.
[13]https://www.malavida.com/en/soft/fakeapp/.
[14]https://github.com/dfaker/df.

Dataset Name	# Real videos		# Fake videos	
	Video	Source	Video	Source
UADFV (2018)[99]	49	YouTube	49	FakeApp
DeepfakeTIMIT (2018)[89]	-	-	620	FaceSwap-GAN
FF++ (2019)[149]	1,000	YouTube	1,000 1,000	FaceSwap[7]. FaceSwap[9].
DeepFakeDetection (2019)[2]	363	28 Attori 16 scene	3068	FaceSwap[9].
Celeb-DF (2019)[100]	890	YouTube	5,639	DeepFake
DFDC Preview (2019)[42]	1,131	66 Attori	4,119	DeepFake
DFDC (2020)[41]	23,654	3,426 Attori	104,500	Multipli
DeeperForensics-1.0 (2020)[77]	50,000	100 Attori	1,000	DeepFake
WildDeepfake (2020)[202]	3,805	Internet	3,509	DeepFake

Table 2.2: *Identity swap dataset.*

to align, colors are blended, and artifacts are removed in a post-processing stage [191].

A relevant dataset for this technique is presented in [140], containing 1,800 images of 150 participants with diverse ethnicities and genders, generated using six different morphing algorithms. Recently, NIST launched the FRVT MORPH evaluation[15], designed to test morphing detection capabilities and robustness against face attacks. Despite ongoing research, the only publicly available dataset is *AMSL Face Morph Image* [124], since datasets involving real identities raise privacy concerns.

2.2.4 Face editing

This manipulation modifies one or more facial attributes, such as eye color, skin tone, hairstyle, gender, aging, or rejuvenation. These techniques are widely used in film and advertising, for example, in cosmetic commercials where models are digitally enhanced by removing imperfections. However, beyond being misleading, such applications can also negatively affect viewers, particularly women, by fostering feelings of inadequacy [148].

Face editing often relies on GANs, as in StarGAN[16] [25], used in the FaceApp application. While it produces impressive results, facial colors sometimes shift

[15]https://pages.nist.gov/frvt/html/frvt_morph.html.
[16]https://github.com/yunjey/stargan

unnaturally. The authors of [129] introduced the Invertible Conditional GAN (IcGAN)[17], which achieves accurate editing but may alter facial identity. An alternative developed around the same time is AttGAN[18] [67]. More recent improvements include STGAN[19] [105], RealGAN [103], and SSCGAN [28].
There are relatively few public datasets generated using these techniques, as researchers often produce custom datasets using the available code. Nevertheless, one widely used dataset is DFFD [33], containing 18,416 real images and 79,960 fake images.

2.2.5 Expression Swap

This manipulation, also known as *reenactment*, allows transferring the expression of a *source* subject onto a *target* one. Beyond facial expression, head and mouth movements can also be transferred, enabling the creation of videos in which someone appears to say things they never actually said. The potential dangers of such technology are evident.

Although expression swapping has existed for years, early methods relied on *computer graphics* to alter frames. With the introduction of DNNs, however, results have become increasingly realistic. Notable projects include DeepFaceLab and DeepFaceLive, both mentioned in the Section 2.2.2. To achieve high-quality results, these systems usually require multiple samples of the *target* face for training. In DeepFaceLab (and in [133]), it is also possible to use just a single target image, although the realism is reduced. Datasets supporting this technique include FF++.

2.2.6 Lip-Sync

Unlike the previous section, these techniques specifically focus on lip movements, conditioned on audio or text. They can therefore be classified into *audio-to-video* and *text-to-video* approaches. Since they are based on speech sounds, recurrent NNs such as *Long Short-Term Memory* (LSTM) are often used, as they are particularly effective with sequential data like words in a sentence. This makes them suitable for modeling lip movements in speech [165]. In [159], whose code is available on GitHub[20], a novel approach based on recurrent and conditional networks is presented, integrating both audio and image features to produce more realistic results, including facial movements and expressions. In [158], an input video is used to generate a 3D model of the *target*, combined with a new video rendering

[17] https://github.com/Guim3/IcGAN
[18] https://github.com/LynnHo/AttGAN-Tensorflow
[19] https://github.com/csmliu/STGAN
[20] https://github.com/susanqq/Talking_Face_Generation

network and dynamic programming method to produce highly realistic videos[21]. Another interesting approach is the Disentangled Audio-Visual System[22] (DAVS) [200], which can be conditioned on both audio and video inputs.

Some studies based on text input include the method presented in [51] and Neural Voice Puppetry [171]. Their respective results can be explored online at https://www.ohadf.com/projects/text-based-editing/ and https://justusthies.github.io/posts/neural-voice-puppetry/.

The FF++ dataset includes this type of manipulation, featuring different kinds of deepfakes such as *NeuralTextures* [172]

The next section explores how these deepfake generation methods have prompted a parallel research field focused on detection and mitigation.

2.3 Deepfake Detectors

The human face serves as a primary biometric identifier and a key channel for non-verbal communication. As shown in the previous section, recent advances in deep image synthesis have enabled the generation of highly realistic synthetic faces and the credible modification of facial attributes in videos. They were presented different methods to fake a video, from face replacement (Face-Swap) to speech modification (Lip-Sync). These manipulations are facilitated by user-friendly applications that can process content in real-time. The rise of this sophisticated deepfake technology has sparked significant alarm in public opinion, driven by fears that it will soon become impossible to distinguish between authentic and manipulated images. This concern extends into critical technical challenges, particularly around the authentication and trustworthiness of digital media. As the line between real and artificial continues to blur, the need for advanced detection techniques becomes paramount to ensure the integrity and reliability of visual information shared across platforms [174, 203, 143, 173, 68].

In the last 20 years, the attention on this topic has kept growing, as reported in [178], even the U.S Department of Defence intended to advance the research on media integrity by launching, through the Defense Advanced Research Projects Agency (DARPA), the large-scale Media Forensic (MediFor) project[23]. Studies focused on detecting manipulated content generated through these techniques are therefore of particular importance, both for safeguarding digital identity, preventing the spread of misinformation on social media, and protecting one's reputation [35]. Such malicious uses can have severe consequences, including financial fraud,

[21] https://wywu.github.io/projects/EBT/EBT.html
[22] https://github.com/Hangz-nju-cuhk/Talking-Face-Generation-DAVS
[23] https://www.darpa.mil/research/programs/media-forensics

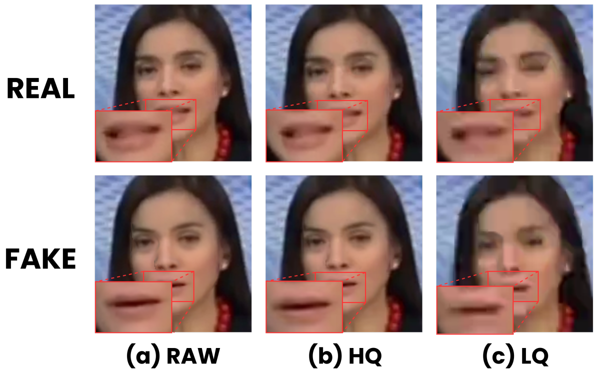


Figure 2.1: *Examples of compressed real and fake images across different compression levels. The top row shows real images, while the bottom row shows fake images. (a) RAW: Original images with no compression, where forged traces in fake images are visible. (b) HQ: High-quality compressed images where forged traces are still noticeable, but compression artifacts are minimal. (c) LQ: Low-quality compressed images with prominent blocking artifacts and significant blurring of forged traces.*

personal and corporate damage, and even geopolitical ramifications [11, 121]. A common threat is therefore represented by the widespread dissemination of such content, which makes the task even more challenging, since multimedia files shared online through social networks are typically compressed before transmission. Compression allows for reduced upload, download, and storage times, but at the same time, it removes information that deepfake detectors have learned during the training process and introduces noise that can mislead the detector. As a matter of fact, compression algorithms, such as MPEG-4 [157] or H.264 [181], introduce artifacts such as blocking, ringing, blurring, or jagged edges, which degrade visual quality and can obscure the features leveraged by detection algorithms. As a result, the accuracy of deepfake detectors significantly decreases in these compressed environments [149, 85, 88]. For these reasons, deepfake detection has advanced significantly in recent years, supported by the release of large-scale datasets and the development of state-of-the-art (SOTA) detection methods.

However, many detectors struggle to handle variations in deepfake typology, compression levels, or resizing. This issue is particularly critical in real-world scenarios, where media is often shared through social platforms and messaging apps that apply compression, degrading data quality (Fig. 2.1) and reducing detection accuracy. While much of the recent literature has focused on improving generalization across different deepfake generation techniques [65], comparatively less attention has been dedicated to robustness against compression artifacts. Yet, compression plays a decisive role in real-world deployment, and its impact on detection reliability is severe. Most existing methods are trained and optimized under clean or lightly compressed conditions, limiting their applicability in practical scenarios. To address this challenge, researchers have explored a range of strategies, primarily within the spatial and frequency domains.

Spatial-based methods analyze image and video data directly in the pixel domain, seeking inconsistencies introduced by manipulation. For example, [120] introduced a CNN followed by a classifier trained on images at intermediate compression levels, thereby improving system resilience. Similarly, Zhou et al. [201] proposed a 3D-ResNext model that extracts features from selected video frames while employing data enhancement techniques during preprocessing to mitigate information loss caused by compression.

In this context, attention mechanisms have proven effective. Attention helps in highlighting significant features within an image or video, crucial under heavy compression, where many details are lost. The authors of [197] introduced a Multi-Attentional Network (MAT) designed to enhance shallow textures and integrate them with deeper semantic features, dynamically adapting attention across facial regions. This adaptability allows the network to maintain focus on important areas that suffer less from compression artifacts, thereby improving detection reliability. The Self-Information Attention (SIA) module, introduced in [164], focuses on high-information regions, enhancing detection accuracy across various models. A different direction was explored in [16], where reconstruction learning, multi-scale graph reasoning, and attention mechanisms were combined to model genuine facial features and detect unknown forgeries, even under highly compressed conditions.

Frequency-based methods were developed under the assumption that compression artifacts manifest more distinctly in the frequency domain. The [112] shows how every GAN leaves its fingerprint, just as a camera is characterized by its PRNU pattern. This fingerprint is kept and exploited in frequency spaces. In [50], a method based on the DCT computed on the entire RGB image is presented, using the result as input to train a CNN. Similarly, in [57], the authors follow the JPEG compression pipeline by splitting the image into 8×8 blocks and then computing the DCT. In [94], it is shown that the AC coefficients, that is, all DCT coefficients except the first one (DC), follow a zero-centred Laplacian distribution. Based on this, the authors extract a unique block from the image using $\beta = \sigma/\sqrt{2}$, where σ is the standard deviation of the AC coefficient distribution. They further demonstrate that this single block value is highly discriminative in detecting fakes, since the β-values of GAN-generated images reveal distinctive fingerprints. In [20], some articles that exploit the Fourier spectrum [44, 46] are cited. These works support the hypothesis that CNN-based generation methods are unable to reproduce the high-frequency distribution of real images, showing, on the other hand, methods that reach a good accuracy with datasets at the state of the art. Furthermore, the authors of [20] reference additional studies proving that, by postprocessing the deep network-generated images or by modifying the last layer of GAN, it is possible to disguise these high-frequency discrepancies. For instance, [136] employed Frequency-Aware Decomposition (FAD) to detect hidden forgery traces and Local Frequency Statistics (LFS) to capture abnormal frequency patterns. Combined in a collaborative framework, these techniques significantly improved detection accuracy, particularly on low-quality images and videos. Liu et al. [104] leveraged phase spectrum analysis to identify upsampling artifacts, enhancing robustness against compression-induced distortions. In [98], the authors combined frequency analysis with metric learning, introducing a Single-Center Loss (SCL) to improve feature separation between real and fake faces, alongside an Adaptive Frequency Feature Generation Module (AFFGM) to capture forgery patterns dynamically.

Relational learning approaches have recently emerged as a promising direction, combining both spatial and frequency cues to model relationships between facial regions. Such methods have proven particularly effective when compression distorts local and global structures. For example, Chen et al. [23] introduced a Multi-scale Patch Similarity Module (MPSM) to capture subtle differences between real and forged facial patches. They also proposed the RGB-Frequency Attention Module (RFAM), which combines RGB and frequency features for stronger local representations. Similarly, [185] presented a framework incorporating a spatio-temporal attention module to capture facial features and a Masked Relation Learner (MRL) to propagate relational information across regions.

Beyond these categories, several works have explicitly addressed the problem of deepfake detection under compression. Zhang et al. [192] proposed a self-supervised decoupling network (SSDN) that leverages compression ratios as self-supervised signals, enabling the model to learn authenticity and compression features jointly. Other strategies include *bleaching* preprocessing, introduced in [97], which simulates compressed images to enhance robustness without retraining. Ref. [24] proposed a learned visibility matrix to focus the model on imperceptible artifacts that remain detectable under compression. The FAMM framework in [102] analyzed facial muscle motions across frames, demonstrating resilience to varying compression levels. Additional approaches, such as the deceptive model in [22], improve generalization and resistance to adversarial attacks, while [198] introduced an anti-counterfeit label mechanism embedding watermarks into facial identity features, making them robust against resizing and compression.

Despite these advancements, significant limitations remain. Most evaluations rely on the FF++ dataset, which includes only two predefined compression levels, 23 and 40 on the Constant Rate Factor (CRF) scale. In contrast, real-world scenarios, such as Facebook Messenger, YouTube, or WhatsApp, employ diverse compression techniques, with ranges that depend heavily on file size and platform requirements. Consequently, two main strategies have emerged: (i) training separate models for each compression level, which is computationally expensive and impractical for deployment, or (ii) training on a single compression level, typically the higher-quality one, to mitigate performance loss. Some works attempt to overcome this by applying data augmentation restricted to the compression levels provided by FF++, thus using them during both training and testing. However, this strategy overlooks real-world scenarios, where social media platforms adopt a wide range of compression levels that extend far beyond those available in the dataset. As a result, the assumed robustness to unseen compression levels does not hold in.

Building on the findings discussed so far, this work introduces two main contributions. The first, which represents the natural continuation of my master's thesis, introduces a frequency-based method leveraging the Discrete Cosine Transform (DCT) to generate a tensor representation of an image across multiple scales and compression levels. The second contribution, currently under review at the time of writing, presents an approach that encodes compression-aware representations directly within the input architecture, by fitting the RGB image directly into the NN, thereby enabling robust detection across a wide range of compression conditions.

2.4 Proposed Research and Contributions

In recent years, as reported in the previous section, after the success of GANs in generating synthetic images, several papers have proposed methods to determine whether an image is real. Some scientific articles present methods based on the Fast Fourier Transform or DCT. However, other papers note that these methods are weak against certain transformations, such as *resize* and *JPEG* or *MPEG* compression in the case of videos [77]. This weakness is very worrying because these transformations are widespread: it's enough to know that, in many photos we receive via messaging apps or on social media, these operations are performed automatically to speed up loading. This section presents two studies: the first, based on the DCT [31], which exploits the traces that manipulations leave in high frequencies; and the second, based on a special Multi-Head module and its scalability in a Multi-Branch Network.

2.4.1 Tensor-Based Deepfake Detection In Scaled And Compressed Images

In this section, the contribution published on the *IEEE International Conference on Image Processing (ICIP 2022)* [31] is shown. The main aim of this study is the design of a deepfake detector able to properly work in a forensic scenario, including common transformations employed in social networks (*e.g.*, resizing and compression) and robust to both background and image size variability. To achieve satisfactory performance also in these challenging conditions, we carefully design a tensor of features provided as input to a custom CNN, properly combining information of the original image at different scaling and compressing levels. Moreover, considering promising results achieved by frequency domain-based approaches [45, 50, 58], the proposed input representation has been built exploiting the DCT domain. Finally, to reduce the mismatch between training and testing data, the contribution of dataset augmentation has also been studied.

2.4.1.1 Feature Extraction

At the basis of this work, there is the conversion of images into the frequency domain, using the DCT. The formula used is shown in Eq. (2.1) in the 2D shape (usually known as DCT Type-II in the various libraries).

$$F[u,v] = \frac{1}{4}C(u)C(v)\left[\sum_{x=0}^{7}\sum_{y=0}^{7} I[x,y]cos(a)cos(b)\right] \tag{2.1}$$

where

$$a = \frac{(2x+1)u\pi}{16}, b = \frac{(2y+1)v\pi}{16}, C(u) = \begin{cases} \frac{1}{\sqrt{2}} & u = 0 \\ 1 & u > 0 \end{cases}, C(v) = \begin{cases} \frac{1}{\sqrt{2}} & v = 0 \\ 1 & v > 0 \end{cases}$$

The DCT is a technique applied to image pixels in the spatial domain to transform them into a frequency domain in which redundancy can be identified [52]. The proposed method consists of splitting an image into different blocks 8×8 pixels with a stride 2×2 and then, for each block, calculating Eq. (2.1) where I is the image. Different from other approaches [58], to increase the amount of collected information, also with small images, overlapping blocks (stride = 2) are considered. This change can be particularly useful when analyzing a small image, as the overlap helps extract more information. The DCT coefficient at position zero is called DC and represents the average of the values in the block. All the other coefficients, namely AC, correspond to specific bands of frequencies. Given all DCT blocks, all DC coefficients can be modelled with a Gaussian distribution while AC coefficients follow a zero-centred Laplacian distribution, as proved in [95]:

$$f(x) = \frac{1}{2\beta}\exp\left(-\frac{|x|}{\beta}\right) \tag{2.2}$$

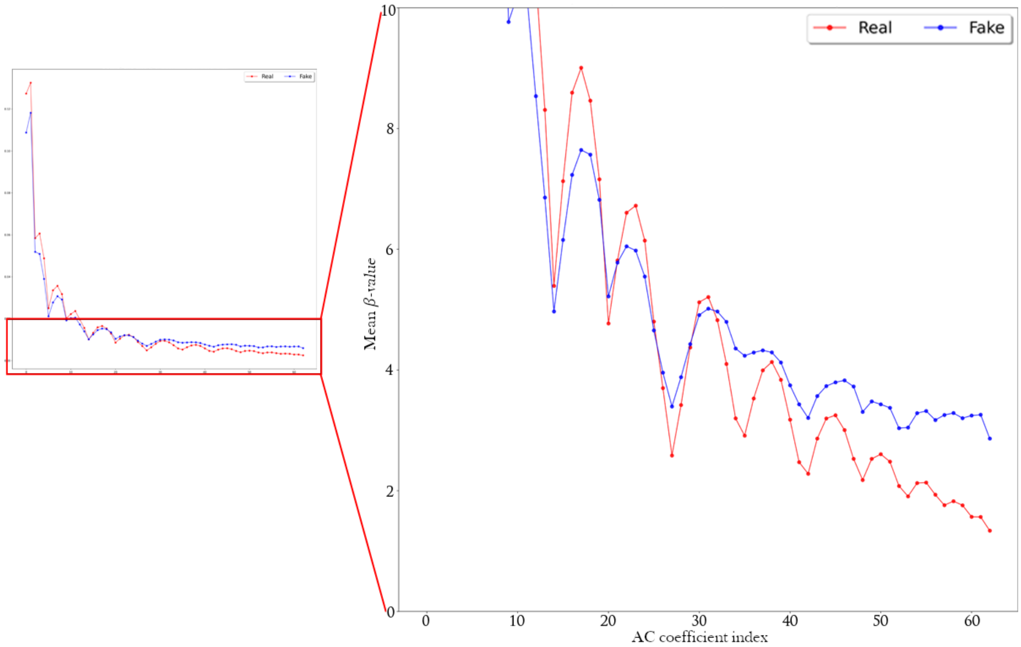


Figure 2.2: *Plot of β-values statistics for fake and real images. The ordinate axis represents the mean of all the image features belonging to the class, while the abscissa axis shows the AC coefficients (zig-zag order).*

With β scale parameter computed by MLE (maximum likelihood estimation), a form solution. Specifically, given AC coefficients $(x^1_{i,j}, x^2_{i,j}, \ldots, x^N_{i,j})$ related to the (i,j) position, $\beta_{i,j}$ has been estimated as follows:

$$\beta_{i,j} = \frac{1}{N_b} \sum_{k=1}^{N} |x^k_{i,j}| \tag{2.3}$$

With N_b number of blocks extracted from I. A beneficial aspect of this method is that the input image can have any dimension: after the pre-processing phase the image will be substituted by a feature vector $\overrightarrow{\beta_I} = \{\beta_{I_1}, \beta_{I_2}, \beta_{I_3} \ldots\}$ with several elements equal to the block dimension minus one: the first element, that is useless because is the average of values in block. Having a fixed-dimensional feature vector, independent of the original image dimensions, is essential because resizing an image to the NN's input size, as we usually do, would lose important clues.

The Fig. 2.2 shows the curves of the mean β-values for fake images versus real images, taken from the OpenForensics dataset [96]. It summarizes the statistical trend of the β-values across images: it is clear that the trends differ, indicating that the values are discriminatory and capable of distinguishing fake images from real ones.

This particular frequency trend is due to the fact that the features in the plot

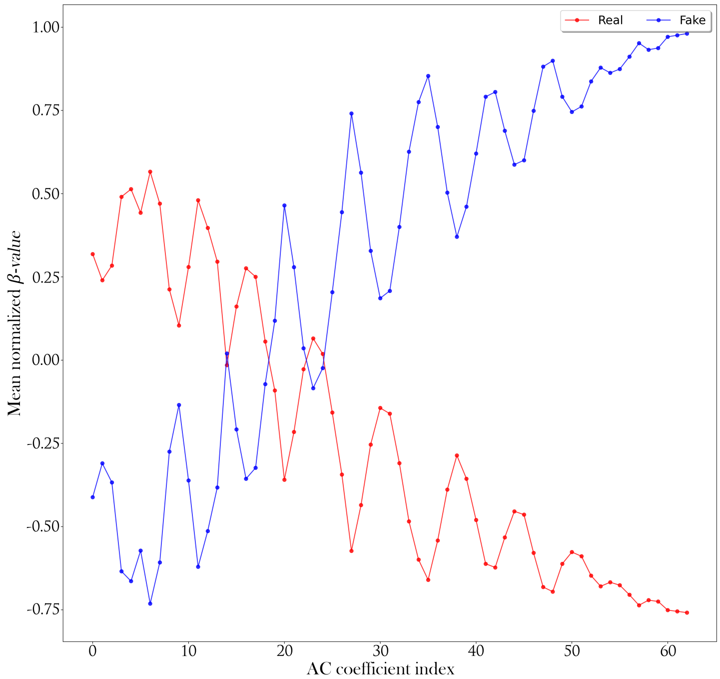


Figure 2.3: *β-values statistics for fake and real images after L_1 and Z_{score} normalization. The ordinate axis represents the mean of all the image features belonging to the class, while the abscissa axis shows the AC coefficients (zig-zag order).*

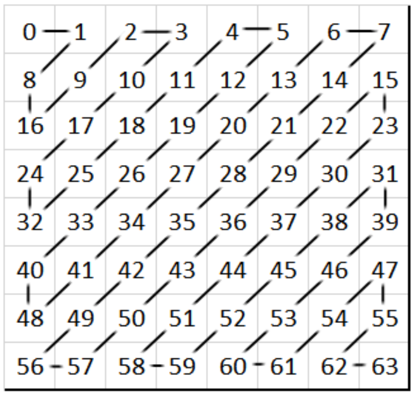


Figure 2.4: *Zig-zag order used after DCT computation.*

in Fig. 2.2 and 2.3 are sorted into a *zig-zag order* (see Fig. 2.4) starting from the top-left element to the bottom-right of the block.

Even more visible is the distance between the two statistical trends of the β-values obtained after the L_1 and the Z_{score} normalization, as shown in Fig. 2.3.

Below the formula of L_1 normalization:

$$L_1 = \left[\frac{\beta_1}{\sum_{i=1}^{63} |\beta_i|}, \frac{\beta_2}{\sum_{i=1}^{63} |\beta_i|}, \frac{\beta_3}{\sum_{i=1}^{63} |\beta_i|}, \ldots, \frac{\beta_{63}}{\sum_{i=1}^{63} |\beta_i|} \right] \tag{2.4}$$

Below the formula Z_{score}:

$$Z_{score} = \frac{(x-u)}{s} \tag{2.5}$$

where u is the mean of the training samples and s is their standard deviation.

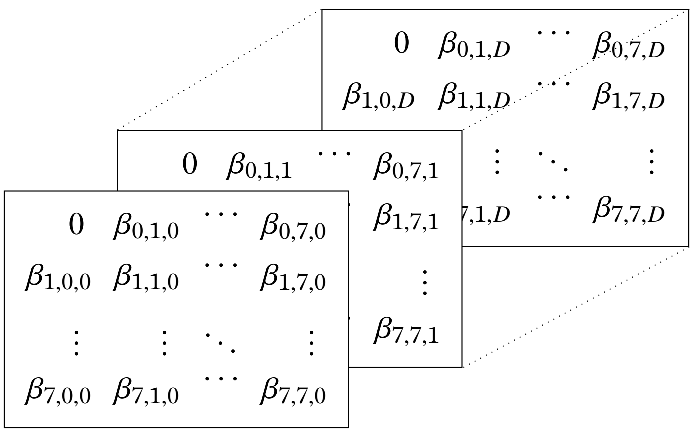
Considering all the AC coefficients, 63 β values can be computed from the input image I. However, differently from [58], to preserve the original positions of DCT coefficients, we collect the features into a matrix B defined as follows:

$$B = \begin{bmatrix} 0 & \beta_{0,1} & \cdots & \beta_{0,7} \\ \beta_{1,0} & \beta_{1,1} & \cdots & \beta_{1,7} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{7,0} & \beta_{7,1} & \cdots & \beta_{7,7} \end{bmatrix} \tag{2.6}$$

Where $B(0,0)$, the value related to DC coefficients that do not follow Laplace distribution [95], has been set to zero.

The representation described in Eq. (2.6) can be further improved considering the typical scenario of social networks and chatting platforms where images and videos are resized and compressed. Such effects have been embedded in the proposed image representation, exploiting information coming from multi-scaling and multi-compression of input data. At first, the original image I is resized N_s times, obtaining a series of images I_s. Compressing then each I_s (N_c times) we obtain $N_s \times N_c$ images $I_{s,c}$. Concatenating then the B matrices Eq. (2.6) obtained from the resized and compressed images $I_{s,c}$ a tensor representation $8 \times 8 \times D$ with $D = N_s \times N_c$ has been derived:

(2.7)



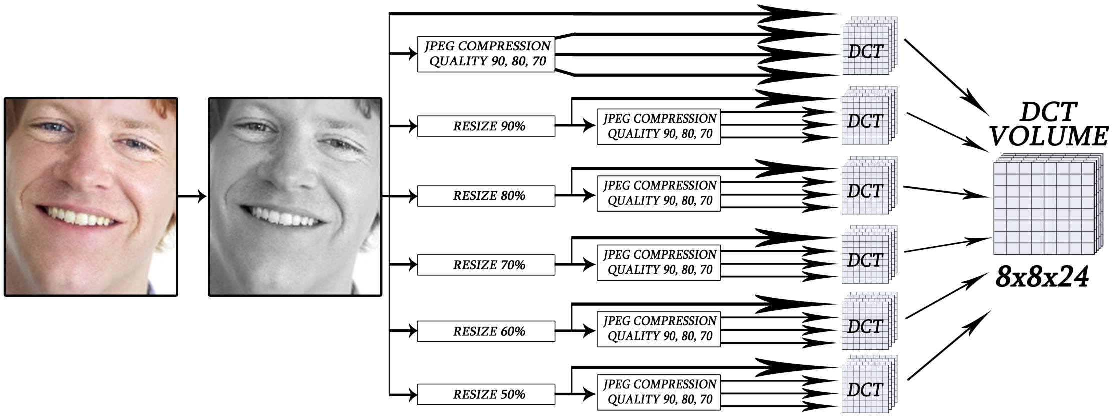


Figure 2.5: *Algorithm's performance that compute the tensor of image and then groups them to make a new image's representation.*

2.4.1.2 Model Architecture

The core of the proposed method consists of exploiting information coming from multi-scaling and multi-compression of input data, according to findings in [113] and [21]. Starting from the typical scenario of social networks/chatting platforms, where images and videos are compressed and resized, we embedded such effects in our feature vector. The rationale behind such an approach is to provide the next module, namely, a deep network, with additional information coming from resized or compressed images, although we cannot know the exact parameters used for such an operation. Therefore, we resized each image H times by a percentage between 90% and 50% and compressed each derived image K times by a JPEG compression between 90 and 70. Thus, the feature vector embeds the β values of the original image, its K compressed versions, and its H resized versions, leading to a feature vector $8 \times 8 \times (K+1) \cdot (H+1)$.

In this study, we focus on the case where $H = 5$ and $K = 3$. We have chosen these particular configurations because they tendentially reflect the typical patterns and constraints encountered in real-world social media platforms. The selected values are representative of how content is often treated and disseminated on such platforms.

Resizing and compressing images was done by the TensorFlow versions of the *Bilinear* resizing method (*Antialias* parameter set to *False*) and the *JPEG compression*, driven by K values of the *quality* parameter, respectively.

To sum up, our feature extraction algorithm, illustrated in Fig. 2.5, consists of:

- computing the DCT of the raw image and the *JPEG compression* version with K quality values, specifically, 90, 80, and 70 ($K = 3$);

- computing H *resized* images at 90%, 80%, 70%, 60%, 50% of the raw version ($H = 5$):
  - for each resizing, computing their DCT and that of *JPEG compression* versions with the K quality values above;
- grouping in unique shape $8 \times 8 \times 24$, the new feature matrix.

The tensor representation (Eq. 2.7) of the original image can then be provided as input to a CNN to perform the final classification. Although simple classifiers [45, 58], or networks with a limited number of layers [50], have been employed to underline the discrimination power of features extracted in the frequency domain, challenging scenarios actually require deep networks to provide satisfactory results (see Section 2.4.1.4). A custom CNN based on the Xception architecture [27] has been designed. As reported in Fig. 2.6, the input volume is processed by the entry flow, then the middle flow is repeated four times, and finally the exit flow is applied. Moreover, to avoid early subsampling of the $8 \times 8 \times D$ input data, pooling layers have been applied only at the end of the network. Finally, a fully connected layer and a dropout layer have been placed before the final logistic regression.

2.4.1.3 Dataset

The dataset for this study was chosen to ensure a thorough analysis of deepfake detection methods. The OpenForensics dataset [96], which forms the basis of this research, contains images sourced from Google Open Images [91] repository. These images are systematically manipulated to create deepfakes, providing a realistic and challenging testbed for forensic analysis.

A key strength of OpenForensics is its diversity, as the dataset covers various indoor and outdoor scenes, including homes, supermarkets, shops, museums, beaches, and mountains. These environments are captured under different lighting and visibility conditions, reflecting real-world complexity and making results more generalizable. The dataset offers facial images at varying scales, ranging from close-ups to distant views, which adds useful variety for detection tasks.

Another important feature is that the images keep their original resolution, since they are not resized. This preserves subtle visual details, which are important for forensic detection. It also allows researchers to test methods that use high-frequency image information. Deepfake changes are made to certain faces, chosen at random in each image. This random selection helps reduce bias and better reflects the unpredictable nature of deepfake creation in real situations. The manipulations use advanced generative models: *InterFaceGAN* [155] or autoencoder *ALAE* [134], making the task more relevant and challenging.

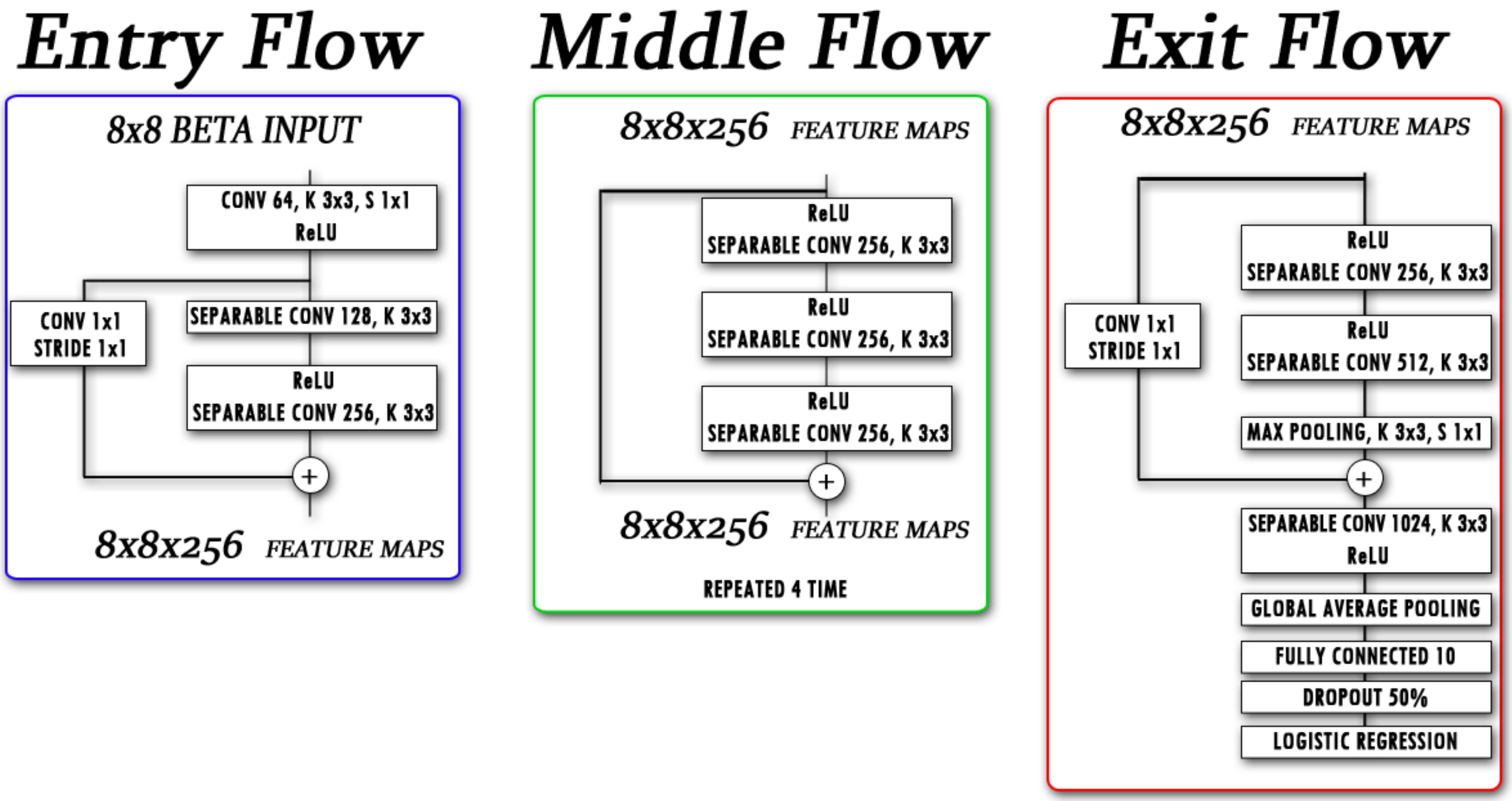


Figure 2.6: *CNN architecture of the custom Xception [27]. The entry flow processes the input volume, then the middle flow is repeated four times, and a final exit flow. Although not reported in the picture, batch normalization follows convolution layers.*

(a) *Error in the yellow bounding box.* **(b)** *Error in the blue bounding box.*

Figure 2.7: *Samples from the OpenForensics dataset with incorrectly sized bounding boxes.*

Table 2.3: *Image distribution in filtered OpenForensics dataset.*

Subset	#Real Faces	#Fake Faces	#Total Faces
Training	81.059	62.689	143.748
Validation	4.742	10.534	15.276
Test Development	20.882	28.564	49.446

The synthetic faces are resized to be positioned on the original faces using Poisson blending [132]. The new identity generated is tested on a classifier trained to recognize fakes. Only if the new image deceives the classifier will it be overlapped on the original one. In summary, OpenForensics was chosen for its technical strengths, diversity, and suitability to the thesis goals: to study and improve deepfake detection in realistic forensic scenarios. The entire picture collection is then split into training, validation, and test sets. For each group, a *JSON* file lists the face detections for each image. The coordinates provided refer to the original image, outlining bounding boxes around the faces. To extract the faces from dataset images, we wrote a script that reads information about the bounding boxes from the *JSON* file. After cutting the bounding boxes, we saved them in *PNG* files, a lossless format. In some cases, only a part of the face was detected, as you can see in Fig. 2.7a and in 2.7b. Some bounding boxes had the minor side smaller than 8 pixels, too few to represent a face. To reduce this problem, we filtered the bounding box dimensions, keeping only those with a smaller side of 30 pixels or more. In Table 2.3, the details of the final dataset used are shown, split from the authors into training (D_{train}), validation (D_{val}), and test (D_{test}) sets.

Table 2.4: *Augmentation details for the training and validation sets.*

Augmentation	Number of copies for each raw img	Resize %	Quality factor
JPG Compression	2	-	Random (70-90)
Resize	2	Random (55-90%)	-

2.4.1.4 Protocol

To test the performance of the proposed solution in the challenging scenario described before, a novel test set including resizing and compression has been built from D_{test}. Specifically, each image of D_{test} has been scaled by a percentage factor $s_f \in \{55\%, 70\%, 85\%\}$ and compressed by a Quality Factor [179] $QF \in \{70, 80, 90\}$. Moreover, to study the effect of both transformations, images obtained by a resizing of 70% followed by a compression with $QF = 80$ have been included. Our goal was to test the performance in case of resizing and/or compression of the images, so all images of the dataset were downscaled in steps of 15%, so down to 85%, 70% and 55% compared to their original size and compressed with a quality factor that is 10 steps lower each time: 90, 80 and 70 compared to their original state. In order to improve the results in the presence of resizing and compression, an augmentation process has also been employed on the training and validation sets as reported in Table 2.4. Note that both resizing and quality factors are picked randomly from uniform distributions.

In the proposed solution, the tensor-based representation defined in (2.7) has been obtained considering six different resizing ($N_s = 6$) uniformly distributed from 50% to 100% and four compressions ($N_c = 4$) with $QF \in \{70, 80, 90, 100\}$. To sum up, an $8 \times 8 \times 24$ tensor is provided as input to the designed custom Xception. The adopted network has been trained using the Adam Solver, with an initial learning rate equal to 0.001, linearly reduced by a factor 0.7 from the third epoch, the Binary Cross Entropy as loss function, and cross-validation to save the best network.

2.4.1.5 Results

To analyze the impact of input representation, dataset augmentation, and the designed CNN architecture, a series of tests has been performed, considering both simple and challenging scenarios, including resizing and compression.

Initially, the accuracy of Matrix-based (Eq. 2.6) and Tensor-based (Eq. 2.7) representations, employing several augmentation strategies (see Table 2.4), was computed.

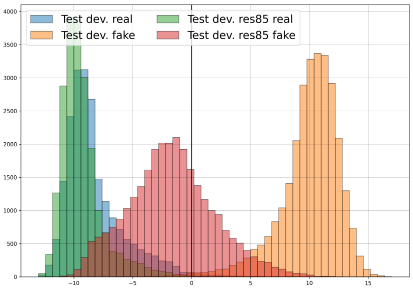
Table 2.5: *Comparison of the results related to the two inputs (Matrix (2.6) and Tensor (2.7)) with and without augmentations employing the custom Xception (see Fig. 2.6). The values in bold represent the highest result for each column.*

Input (Xception)	AUG.	Test Development Accuracy %									
		Raw	Resize %			JPEG Comp. (QF)			Res70 + JPEG80	AVG	SD
			85	70	55	90	80	70			
Matrix (2.6)	no	99.2	59.1	74.1	77.5	57.5	44.1	42.2	42.4	62.0	20.3
Tensor (2.7)	no	**99.4**	72.9	89.4	79.3	54.1	42.2	42.3	42.3	65.2	23.0
Matrix (2.6)	Resize	98.5	98.2	93.1	78.5	73.3	50.2	43.6	44.9	72.5	23.6
Tensor (2.7)	Resize	99.3	**98.5**	94.9	89.0	45.2	42.5	42.4	42.9	69.3	28.1
Matrix (2.6)	JPEG	98.2	87.0	89.8	79.2	97.2	94.2	91.8	78.5	89.5	7.5
Tensor (2.7)	JPEG	99.2	75.0	90.4	81.1	97.5	**96.9**	**94.3**	82.3	89.6	9.0
Matrix (2.6)	Both	98.3	97.9	93.7	84.3	97.0	94.6	92.1	79.0	92.1	6.9
Tensor (2.7)	Both	**99.4**	98.1	**95.2**	**89.3**	**98.0**	**96.9**	93.9	**83.8**	**94.3**	**5.3**

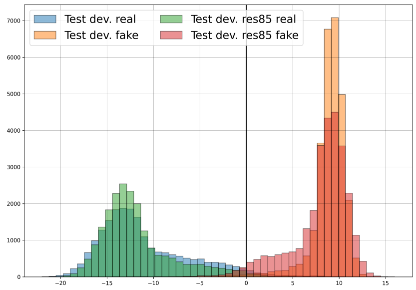
As reported in Table 2.5, the former representation without any augmentation achieves satisfactory results on the original raw dataset (99.2% accuracy), but its performance degrades quickly when the test set includes resized and compressed images. The proposed Tensor-based representation (Eq. 2.7), even without any augmentation, is able to reach higher accuracy on the resized images but not on the compressed ones. Both representations improve their performance considerably when trained on the specific type of manipulation. Finally, considering an augmentation strategy that includes both resizing and compression, a satisfactory degree of robustness with respect to these transformations is achieved.

The Tensor-based representation attained the best performance employing the augmentation strategy, including both transformations. It is also interesting to note that in the *Res70 + Jpeg80* column, which represents images first resized to 70% and then compressed with a quality factor of 80 (not present in any augmentation), the results are better for the models considering the JPEG augmentation strategy.

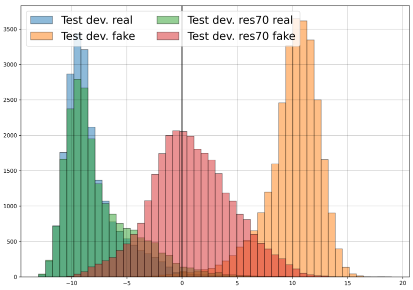
These results confirm our initial hypothesis: it is indeed possible to improve the performance of a deepfake detection system in a real scenario by combining information of the original image at different scales and compressing levels and by exploiting dataset augmentation techniques.
Figures 2.8 and 2.9 show the score distributions computed by the custom Xception right before the softmax function in the case of Matrix-based (2.6) and Tensor-based (2.7) representations with (second column) and without (first column) augmentation. We report the worst cases of images resized by 85% and 70% of their original dimensions and compressed with quality factors of 80 and 70, respectively.



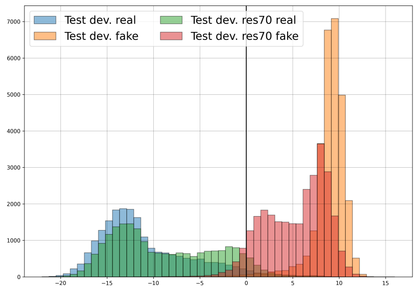
(a) *Resize 85%*



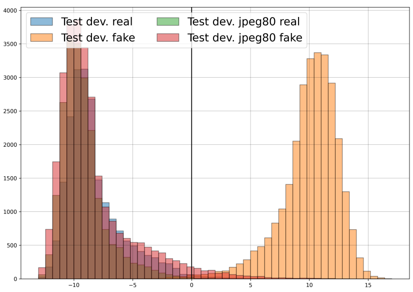
(b) *Resize 85% (augmented)*



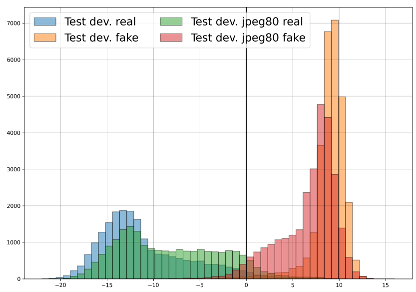
(c) *Resize 70%*



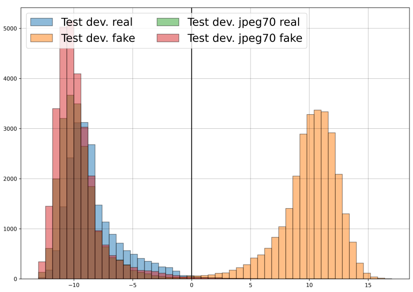
(d) *Resize 70% (augmented)*



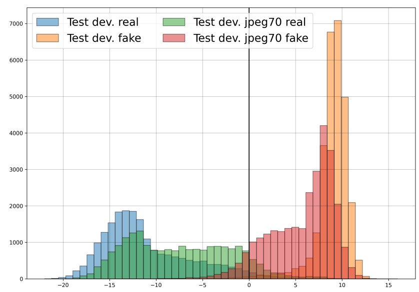
(e) *Compression QF 80*



(f) *Compression QF 80 (augmented)*

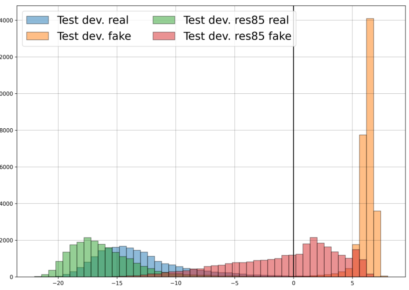


(g) *Compression QF 70*

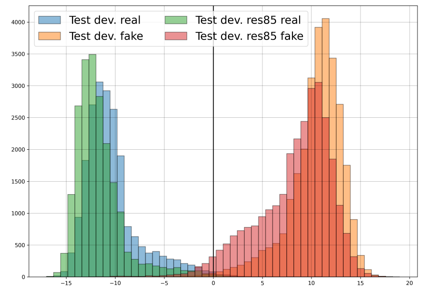


(h) *Compression QF 70 (augmented)*

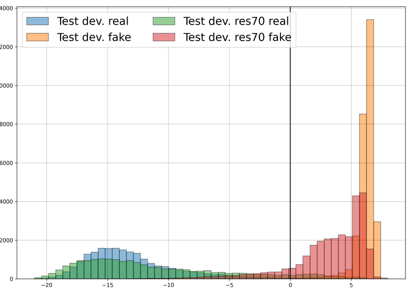
Figure 2.8: *Score distributions of the network right before the softmax function, considering **Matrix** representation, without augmentation in the first column and employing the augmentation in the second column*



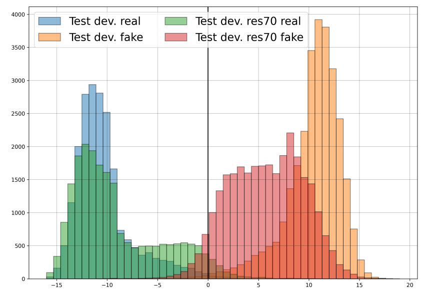
(a) *Resize 85%*



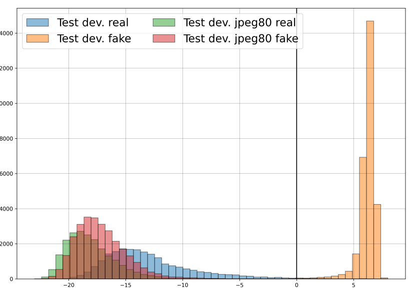
(b) *Resize 85% (augmented)*



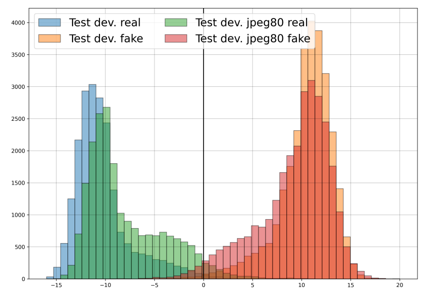
(c) *Resize 70%*



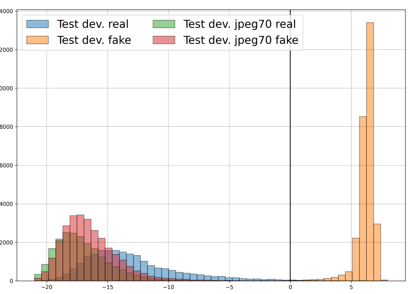
(d) *Resize 70% (augmented)*



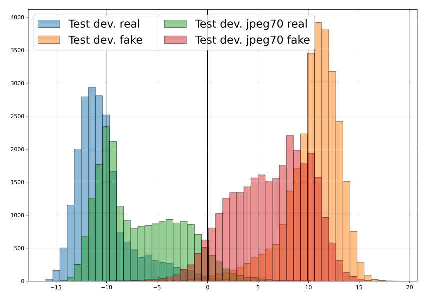
(e) *Compression QF 80*



(f) *Compression QF 80 (augmented)*



(g) *Compression QF 70*



(h) *Compression QF 70 (augmented)*

Figure 2.9: *Score distributions of the network right before the softmax function, considering **Tensor** representation, without augmentation in the first column and employing the augmentation in the second column.*

Table 2.6: *Comparison of the proposed method with state-of-the-art DCT methods.*

Method	Test Development Accuracy %									
	Raw	Resize %			JPEG Comp. (QF)			Res70 + JPEG80	AVG	SD
		85	70	55	90	80	70			
CTF [58]	89.8	88.3	87.9	76.6	84.0	79.7	81.0	70.9	82.3	6.5
Rid.-Reg. [50]	90.5	90.0	88.4	81.0	90.1	89.1	85.3	76.3	86.3	**5.2**
Matrix (Logistic Regression)	94.1	93.2	90.5	77.8	95.4	89.9	85.4	69.3	87.0	9.1
Tensor (Logistic Regression)	95.8	93.1	91.6	80.1	95.3	92.7	88.4	76.5	89.2	7.2
Matrix (Xception)	98.3	97.9	93.7	84.3	97.0	94.6	92.1	79.0	92.1	6.9
Tensor (Xception)	**99.4**	**98.1**	**95.2**	**89.3**	**98.0**	**96.9**	**93.9**	**83.8**	**94.3**	5.3

Fig. 2.8 represents the scores related to the Matrix representation (Eq. 2.6) when tested on the raw data and resized images or compressed ones. We observe that in raw data, the real and fake distributions are well separated, enabling the system to achieve satisfactory performance. However, for resized and compressed data, fake distributions shift to the left of the plot. This fact is less pronounced when considering the Tensor-based representation (Eq. 2.7) (Fig. 2.9), where this shift has been dampened. Finally, as reported in the second column of Fig. 2.8 for Matrix and in Fig. 2.9 for Tensor, the augmentation process enables the correct separation of the fake and real distributions, thus confirming our hypothesis. The proposed solution was also compared with state-of-the-art approaches based on DCT [58, 50]. In particular, in [58] the authors take a gray-scale image, subdivide it into non-overlapping 8×8 blocks, compute DCT, and obtain a feature vector consisting of 63 β-values (DC term is not considered). A Gradient Boosting classifier with the following hyperparameters is then employed: *number-of-estimators* = 100, *learning-rate* = 0.6, *max depth* = 2. In [50], on the other hand, DCT is computed without performing any splitting on the RGB image, after resizing it to 128×128. At first, input data are normalized in the $[-1, 1]$ range before the DCT computation. After these steps, the log scale and the Z-score normalization are applied to DCT data. At the end, a CNN has been used as a classifier. To remark on the expressive power of the proposed representation, two versions have also been considered. The Tensor-based representation has been combined with both a simple (logistic regression) and a deep network (custom Xception). Moreover, all the compared approaches have been trained with the augmented OpenForensics dataset, including resizing and compression (see Table 2.4).

As reported in Table 2.6, the proposed solution outperforms state-of-the-art approaches by a large margin in all the considered scenarios. Although the Matrix-based representation with Logistic Regression achieved state-of-the-art results, the Tensor-based version further improves this outcome. Moreover, the deep network architecture (*i.e.*, the custom Xception) can exploit further both representations in the presence of challenging conditions, making the Tensor-based version the best.

2.4.1.6 Conclusions

The primary contribution of this work is a Tensor-based representation of deepfake images, achieved by combining features from the original image at various levels of scaling and compression. The Tensor-based representation improves the performance of the single-matrix version. Its expressive power, superior to that of similar models when used to train a lighter architecture (i.e., logistic regression), was enhanced by a custom DNN.

In summary, previously unseen, randomly scaled, and compressed deepfakes were more effectively detected by incorporating information about the scaling and compression operations on the input image into the feature set. In future work, it may be interesting to evaluate the possibility of extending the tensor with other image features to maintain the high level of accuracy of deepfake detectors even in the presence of strong transformations.

2.4.2 Robust Deepfake Detection in Compressed Videos with Scalable Network Strategies

Extracting frequency information from images can help detect deepfakes, but it often requires a lot of computing power and may not be practical. To overcome this, the current section explores an alternative approach: teaching the model to recognize compression conditions by adding compression estimation as an additional task. This method addresses three main challenges in real-world detection:

- handling a wide spectrum of compression levels with a single model;
- ensuring efficiency on resource-constrained devices;
- preserving detection accuracy under varying quality conditions.

The first element of the proposed framework is a video compression augmentation pipeline designed to generate multiple compressed versions of each video, systematically covering the compression spectrum and approximating the transformations typically applied by social media platforms [111]. This procedure aims to expose the model to a broad distribution of compression artifacts during training.

Building upon this foundation, two complementary architectural designs are investigated: the Multi-Head Network (MHN) and the Multi-Branch Network (MBN). The MHN extends a standard backbone by restructuring its terminal layer into two specialized heads: one for deepfake classification and another for compression-level estimation. This joint formulation enables the network to simultaneously learn to discriminate between authentic and manipulated content while explicitly modeling the degree of compression, thereby achieving greater robustness across compression domains.

The MBN further develops this concept through a modular design composed of multiple parallel branches, each instantiated as an independent MHN. Unlike conventional deep architectures, which increase capacity by stacking additional layers (often at the cost of computational overhead), the MBN introduces horizontal scaling. Each branch is trained on a bootstrap-resampled subset of the training data, promoting representational diversity. In this way, the MBN achieves a favorable balance between efficiency and performance, leveraging multiple lightweight MHN units rather than relying on a single deep and computationally expensive model.

2.4.2.1 Method

This section introduces a framework designed to improve the detection of deepfakes, focusing especially on facial manipulations found in compressed digital media. One of the main challenges addressed here is the variation in detection performance that occurs at different levels of video compression, a common issue in practical scenarios. To address this, the method aims to reduce the loss of accuracy that can result from diverse compression settings frequently seen in real-world content. The general structure of the framework is shown in Fig. 2.10. The process begins with data augmentation, where each input video is compressed multiple times, using randomly selected compression levels within a specified range. After this, a preprocessing step locates and crops faces directly from these compressed videos. The resulting face images are then used for training. These images are fed into several separate MHN (see Section 2.4.2.3), each tasked with both deepfake classification and recognizing the compression level. The results from these networks are then combined using a Multi-Branch Network (MBN) (Section 2.4.2.4), which allows the system to take advantage of multiple models working together, thereby making detection more reliable. The following sections describe each part of the proposed detection system in detail.

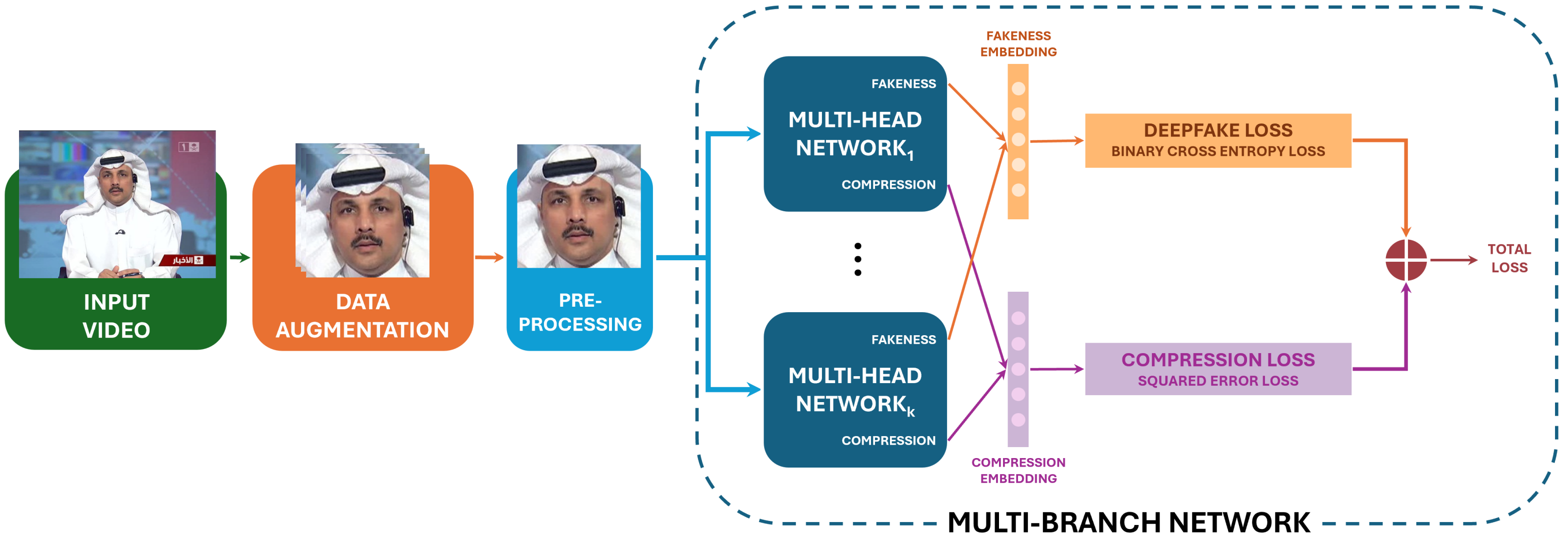


Figure 2.10: *Overview of the proposed deepfake detection framework. Each video is first augmented at multiple compression levels, then face regions are cropped and processed by parallel MHNs for joint deepfake detection and compression estimation. The MHNs form the MBN.*

2.4.2.2 Compression-based Data Augmentation

When working with deepfake images, it is clear that using data augmentation can make detection systems stronger and more adaptable, even when facing different datasets or manipulation methods [180, 12]. Often, this means tweaking the training images by compressing, rotating, scaling, or flipping them. Still, most of these approaches have not been fine-tuned to specifically handle compression artifacts. Furthermore, transitioning from images to videos presents additional challenges that require careful consideration.

The compression of a single image is relatively straightforward: you just apply the compression function to that image. However, with videos, things become trickier, as you need to consider not just each frame on its own, but also how frames relate to each other over time [109]. While image augmentation methods work well for pictures, they don't really capture these extra complexities found in videos. For this reason, we propose a novel augmentation strategy for deepfake video detection that ensures the model sees a wide range of compression levels during training.

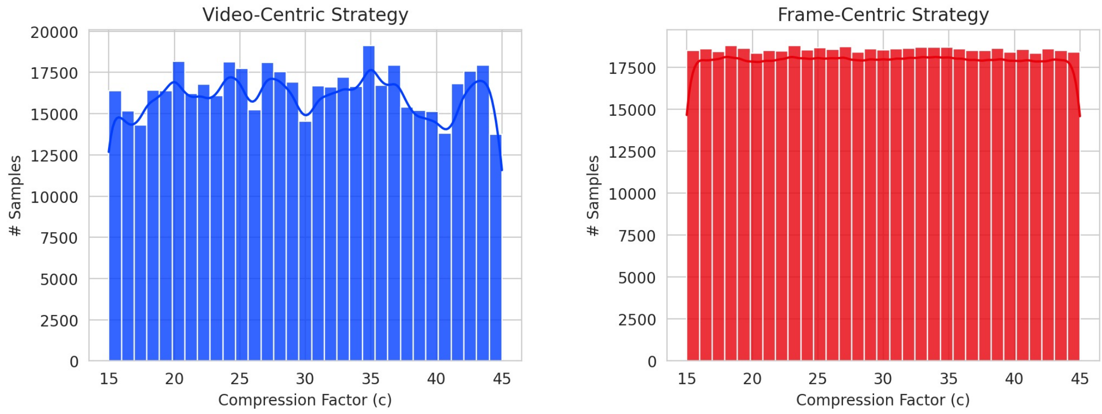
To formalize the proposed augmentation strategy, let's define N_{tot} as the total number of frames used for training. We can write this as:

$$N_{tot} = N_v \cdot N_f \cdot N_q \tag{2.8}$$

In this formula, N_v is the number of videos, N_f is how many frames are taken from each video, and N_q is the number of compression levels picked at random from a set interval $\mathcal{I} = [c_{min}, c_{max}] \cap \mathbb{Z}$, where c stands for a general compression parameter. For each original video v, a set of compressed versions $\mathcal{V}$ is generated, which is defined as:

$$\mathcal{V} = \{v_c \mid c \in \mathcal{I}\} \tag{2.9}$$

Here, v_c just means the version of video v that's been compressed using parameter c. With all these compressed videos in $\mathcal{V}$, there are two main ways to pick N_q compression levels for the training set: a *video-centric* (*vc*) method and a *frame-centric* (*fc*) method. The main difference lies in how these compression levels are distributed across the videos and frames. In the *vc* strategy, for each video v, N_q compressed versions are randomly selected from $\mathcal{V}$. Subsequently, N_f frames are randomly chosen from these versions to constitute N_{tot} training frames.



(a) **(b)**

Figure 2.11: *Distribution of sampled frames per compression level* $c \in \mathcal{I} = [15, 45]$ *for the vc (a) and fc (b) augmentation strategies with* $N_q = 2$ *and* $N_f = 100$*. Compression is performed using the H.264 codec, where* c *denotes the CRF.*

As illustrated in Fig. 2.11a, this approach can result in certain compression levels being disproportionately represented due to the randomness of the selection process. For instance, if $\mathcal{I} = [15, 45]$, $N_q = 2$, and $N_f = 100$, each video is associated with two randomly selected CRF values (such as 18 and 33), and the 100 frames are drawn exclusively from these two versions. Consequently, some compression levels may be overrepresented or underrepresented in the training dataset. In contrast, the *fc* method begins by selecting N_f frames from each original video v. For each frame, N_q compressed versions are randomly chosen from $\mathcal{V}$. This approach also yields N_{tot} frames. The key distinction is that each frame is associated with multiple compression levels, which generally produces a more uniform distribution of compression levels in the dataset. As shown in Fig. 2.11b, the resulting distribution is flatter and more balanced than in the vc case, even when using the same settings.

Both strategies result in training sets with the same number of frames, ensuring no difference in dataset size or computational requirements during training. The total frame count is consistent with those used in previous studies. Notably, this approach enables exploration of a broad range of compression levels rather than limiting analysis to a single setting. Section 2.4.2.8.2 will compare these two strategies and demonstrate their respective impacts on deepfake detection performance. Additionally, an extra component was incorporated into the neural network to estimate the compression level for each input; this feature will be detailed in the following section.

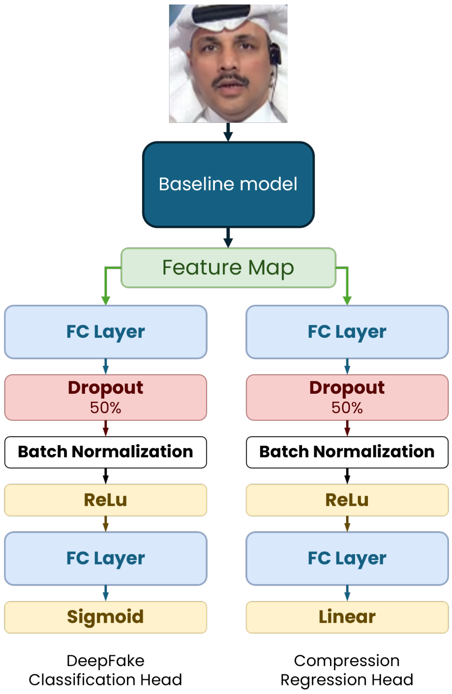


Figure 2.12: *MHN Architecture. From a shared feature map generated by the baseline model, the MHN architecture diverges into two parallel heads dedicated to deepfake classification and compression level estimation tasks.*

2.4.2.3 Multi-Head Network (MHN)

To integrate compression-related information into the learning process, we introduce the Multi-Head Network (MHN), a two-branch architecture designed to address deepfake classification and compression level estimation jointly. As illustrated in Fig. 2.12, the MHN extends a standard backbone by appending two parallel output heads: a classification head that distinguishes between deepfake and authentic samples, and a regression head that estimates the level of compression applied to the input frame. This dual-head design is inspired by multi-task and multi-label learning, which are well-established concepts in machine learning. Multi-task learning scenarios often utilize multi-head models where each head is responsible for a different task, improving the overall performance by leveraging shared representations [196]. Similarly, in multi-label classification, different heads can handle distinct labels or categories, optimizing the model's capacity to learn complex relationships within the data [168].

Based on these principles, we can view the problem of deepfake detection under compression as a specific case of multi-task learning. Each head follows a fixed architectural structure while allowing flexibility in its internal configuration to adapt to different backbone architectures or computational constraints. In our configuration, we employed the following components:

- *Fully Connected (FC) layer*: the first dense layer projects the features extracted by the backbone into a lower-dimensional representation. The number of units can be adjusted based on the model capacity and dataset characteristics, and is treated as a design parameter within the architecture.

- *50% Dropout layer*: randomly set half of the neurons to zero during each training step. A dropout rate of 0.5 is commonly used as it balances retaining enough information for learning while effectively mitigating overfitting, as suggested in [161].

- *Batch normalization*: included to stabilize the learning process and accelerate convergence;

- *ReLU activation function*: a widely adopted non-linearity function that enables efficient gradient propagation and stable training;

- *FC with a single neuron*: the classification branch uses a sigmoid activation function to output a probability score for the binary classification task (real vs. fake), while the regression branch uses a linear activation function to output a continuous value representing the compression level of the input image.

To construct a MHN from a pre-trained model, depending on the network, it may be necessary to add a layer to connect to the fully connected layers of the MHN. We used a Global Average Pooling 2D layer for this purpose. Otherwise, removing only the final layer suffices to connect to the MH module without additional layers.

The overall loss function in the MHN is a weighted sum of the losses from the classification and regression tasks. The classification branch uses Binary Cross Entropy (BCE) loss. For a batch of n samples, it is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \tag{2.10}$$

where $y_i \in 0,1$ is the true binary label and $\hat{y}_i \in [0,1]$ is the predicted probability for the i-th sample.

The regression branch uses Squared Error (SE) loss. To maintain consistency with the BCE loss, we define it as the mean over the batch:

$$\mathcal{L}_{SE} = \frac{1}{n}\sum_{i=1}^{n}(r_i - \hat{r}_i)^2 \tag{2.11}$$

where r_i is the true compression level and $\hat{r}_i$ is the predicted compression level for the i-th sample.

The total loss function, combining these components, is then given by:

$$\mathcal{L}_{tot} = w_{BCE} \cdot \mathcal{L}_{BCE} + w_{SE} \cdot \mathcal{L}_{SE} \tag{2.12}$$

Equal weights were assigned to the classification and compression detection tasks ($w_{BCE} = w_{SE} = 0.5$) to balance their influence during training.

2.4.2.4 Multi-Branch Network (MBN)

Traditional methods often try to improve performance by making networks deeper, but this can raise issues with efficiency and real-world use. Deeper models add more layers in a single sequence, which may not work well for on-device deepfake detection, especially when resources are limited or inputs are low quality, like compressed videos from social media. The MHN helps classify these videos without adding many parameters, but it still requires complex backbone architectures that demand significant computation.

To address these challenges, we introduce the Multi-Branch Network (MBN), a modular framework with B parallel branches, each acting as its own MHN. Instead of making the model deeper, MBN spreads its capacity across several branches, which allows for more flexibility in managing complexity and performance. This design is inspired by ensemble learning, where combining the results of different classifiers often leads to better generalization. A fundamental element underpinning the effectiveness of ensemble methods is the diversity among constituent classifiers, which enables the aggregate model to exploit complementary information and more effectively capture data variability [184]. To foster diversity within the MBN, this study employs the ***bootstrap aggregating*** (bagging) technique, which is rooted in the bootstrap resampling approach pioneered by Efron et al. [47] and subsequently formalized in the context of ensemble learning by Breiman [14].

The bootstrap is a non-parametric resampling method designed to estimate the sampling distribution of a statistic. Consider a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where N denotes the number of samples, $\mathbf{x}_i$ represents the input feature vector, and y_i is the associated label. For each bootstrap replicate $\mathcal{D}^{(b)}$, sampling is performed N times with replacement from $\mathcal{D}$, for $b = 1, \ldots, B$, with B being the total number of replicates. An important characteristic of the bootstrap method is that, owing to the use of sampling with replacement, each replicate typically contains approximately 63.2% of the unique instances from the original dataset. Mathematically, the probability that a particular sample is excluded from a replicate is $(1 - \frac{1}{N})^N$, which, as N approaches infinity, converges to $e^{-1} \approx 0.368$. Consequently, the likelihood that the sample appears at least once in a replicate equals $1 - e^{-1} \approx 0.632$.

In the MBN, each bootstrap replicate $\mathcal{D}^{(b)}$is assigned to its own branch. For consistency, all B branches use the same MHN design, though the framework can also support different architectures if needed. While adding more branches increases the total number of parameters, the overall complexity can be controlled during design. By carefully choosing the number of branches and their architectures, the model can be tailored to fit different resource limits. The main goal is to show that this method of adding diversity leads to better results than single-branch models, even those with more parameters.

From an architectural perspective, concatenation is performed at the two penultimate layers within each MHN. As depicted in Fig. 2.13, the terminal layers of the heads are omitted, with the fully connected (FC) layers responsible for deepfake classification concatenated on one side, and those associated with compression level regression concatenated on the other. Subsequently, these concatenated layers are processed by a *Dropout* layer with a dropout rate of 0.5, which promotes engagement across all neurons and branches. Finally, an FC layer, in conjunction with a sigmoid activation function, is employed for the classification (fakeness) branch, whereas an FC layer, coupled with a linear activation function, is utilized for the compression regression branch.

2.4.2.5 Datasets

In the first contribution (2.4.1), the OpenForensics [96] dataset was selected because it represents one of the most comprehensive benchmarks, including modern manipulation techniques and offering high variability in acquisition conditions (*e.g.*, indoor and outdoor scenes, day and night lighting). Since that study examined how JPEG compression affects manipulated images, OpenForensics was an appropriate choice thanks to its variety of visual scenes and manipulation techniques.

In contrast, the present study focuses on video compression and its impact on deepfake detection performance. For this reason, two well-established video datasets were selected: FaceForensics++ (FF++) [149] and the Celeb-DF (v2) [101].

FF++ is a widely used dataset for evaluating deepfake detection systems. It comprises 1,000 real and 4,000 fake videos, created using four different face manipulation techniques: FaceSwap, Face2Face, DeepFakes, and NeuralTextures. The first two are graphics-based, while the remaining ones are learning-based. The videos in FF++ are sourced from YouTube and feature various individuals in various scenarios. The authors selected only those in which the target face tends to maintain a frontal position to avoid artifacts caused by the attacks. Regarding ethnic diversity, the dataset shows a reasonable variety, though not balanced, while it is well-balanced in terms of gender.

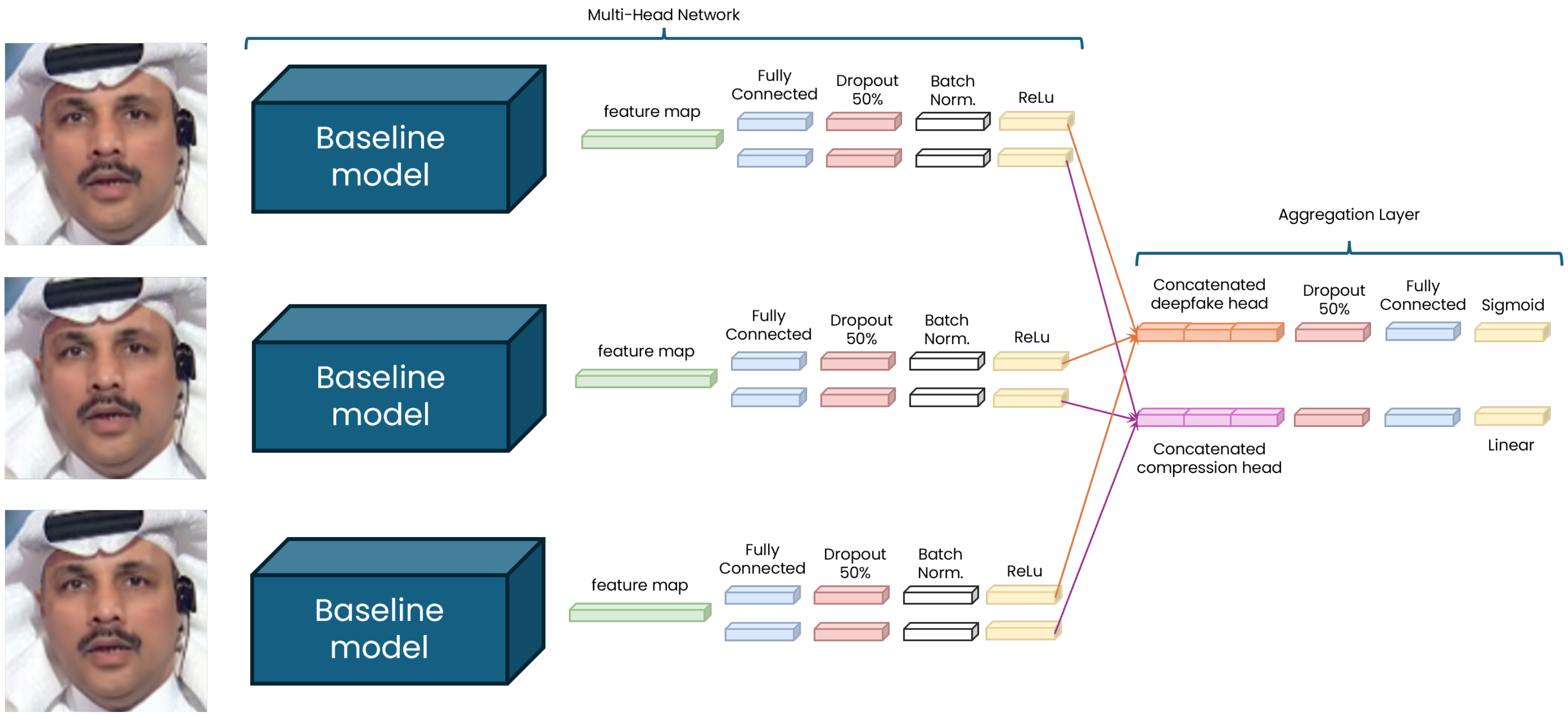


Figure 2.13: *MBN Architecture. Each branch represents a replicated MHN, independently trained to introduce diversity. The final layers of each branch are concatenated to aggregate features, providing a comprehensive representation for the subsequent classification and regression tasks.*

The dataset is split into three versions based on compression levels using the H.264 codec: raw (videos uncompressed), c23 (HQ, high-quality), and c40 (LQ, low-quality). The c23 and c40 versions represent videos compressed with Constant Rate Factors (CRF) of 23 and 40, respectively. Additionally, they provide videos showing only the manipulated mask in cases of fake videos. The diversity of manipulation techniques and compression levels in this dataset establishes a comprehensive benchmark for evaluating the generalization and robustness of deepfake detection models.

Celeb-DF (v2) is a challenging dataset for deepfake detection, with 890 real and 5,639 fake videos. The real videos feature 59 celebrities sourced from YouTube, cover diverse facial expressions, poses, and lighting conditions. The fake videos were generated using an improved DeepFake synthesis process that enhances visual quality and realism, making the dataset particularly valuable for evaluating the performance of detectors in real-world scenarios. Since this dataset does not provide any compressed versions, we applied the same compression levels as used in FF++ (CRF values of 23 and 40) to Celeb-DF (v2) to ensure a direct comparison between the two datasets.

2.4.2.6 Protocol

This section describes the experimental protocol used to evaluate the deepfake detection framework. The FF++ dataset was split into training, validation, and test sets at 720:140:140, consistent with the original settings in [149]. The Celeb-DF (v2) dataset was split into training and test sets according to the official guidelines.

A uniform sample of $N_f = 100$ frames was extracted from each video in the datasets to comprehensively assess model performance. Compression was applied using the H.264 codec, with compression levels sampled from the interval $\mathcal{I} = [15, 45]$, where $c_{min} = 15$ indicates light compression and $c_{max} = 45$ indicates strong compression. This range was selected to represent the variety of compression levels typically encountered in real-world scenarios, where lossy compression is used to reduce bitrate and storage requirements. The complete set of compressed versions $\mathcal{V}$ was generated as defined in Eq. (2.9). Of the two augmentation strategies introduced in Subsection 2.4.2.2, the *fc* strategy was adopted, as it yielded superior results (see Subsection 2.4.2.8.2 for comparative analysis). The number of compression levels used in the augmentation process was fixed at $N_q = 2$ to ensure each raw frame was associated with at least two compression levels and to provide variability across training samples. Consequently, the total number of training frames used was $N_{tot} = N_v \cdot N_f \cdot (1 + N_q) = N_v \cdot 300$.

Subsequently, we employed the Multi-task Cascaded Convolutional Networks (MTCNN) [188] to detect faces in the scene. Given the presence of multiple faces in

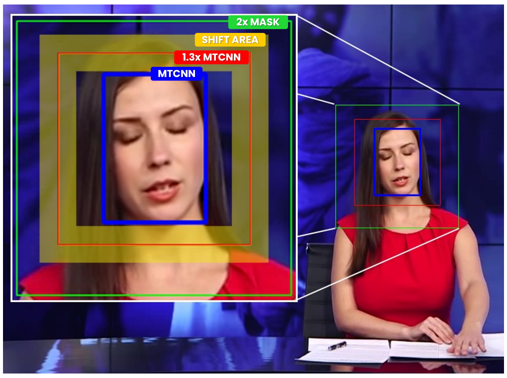


Figure 2.14: *Face cropping: the green box represents the scaled mask region (x2), the blue box shows the initial MTCNN [188] detection, the red box is the extended bounding box (x1.3), and the yellow shaded area indicates potential shifts up to 10% of its side length.*

certain videos, particularly those in the FF++ dataset, it was crucial to ensure that the extracted face was the manipulated one. In some cases, errors were found in which the manipulation was applied to background regions rather than the target face. For proper training, our algorithm selects only the bounding box within the mask area provided by the dataset, applying a conservative approach by scaling it by a factor of 2 to ensure the entire face region is captured. Once the correct bounding box was obtained, we followed the methodology of [149], extending it by a factor of 1.3 while keeping the center fixed and maintaining a square shape. This enlargement was necessary to include slightly external facial features, such as the neck, where manipulations often occur during attacks, such as lip-sync. Moreover, to ensure that the network utilizes all available pixels effectively, we applied a random shift, either negative or positive, to the bounding box coordinates by a variable amount up to 10% of the bounding box side length. This additional augmentation step occurs with a 75% probability on each frame. Finally, the cropped face was resized to 224x224x3 pixels and normalized to the range [-1, 1] for network input. The entire process is illustrated in Fig. 2.14.

We evaluated several baseline network architectures, chosen for their optimal balance between performance and computational efficiency:

- *MobileNetV3Small* [71]: With its low parameter count (~1M), it is suitable for mobile and embedded applications, providing a lightweight yet effective option for real-time deepfake detection.

- *MobileNetV3Large* [71]: This variant offers increased capacity and complexity, making it more suitable for scenarios requiring higher accuracy without significant computational overhead. It has approximately 3M parameters.

- *EfficientNetV2B0* [167]: Known for its balance between accuracy and efficiency, EfficientNetV2B0, with approximately 6M parameters, provides a robust framework for deepfake detection.

- *Xception* [26]: Renowned for its high accuracy in image classification tasks, Xception serves as a common benchmark in deepfake detection studies, offering a comprehensive evaluation of our approach's effectiveness, albeit at the cost of significantly higher computational demands (~21M parameters).

We incorporated our proposed Multi-Head (MH) and Multi-Branch (MB) enhancements into these baseline architectures. Notably, these modifications are applicable to any backbone. The selected baselines exemplify the efficiency and scalability of our approach across a representative set of architectures, particularly with respect to computational cost and parameter count.

2.4.2.7 Implementation details

The MHN is generated by keeping the feature extraction part of the baseline model (pre-trained with ImageNet), adding a Global Average Pooling layer, and two heads. Each head includes a series of layers as described in Subsection 2.4.2.3, and the parameters initialization used are the default in Keras. The fully connected layers are initialized with Glorot uniform initialization for kernels [59] and zero for biases; no seed is set in the dropout layer; batch normalization has a momentum of 0.99 and epsilon of 0.001. The MBN training process consists of several stages. Initially, each branch is trained separately as an individual MH model using stratified bootstrap resampling of the training set. Stratification is applied solely to the binary class label (real versus fake), and no constraints are placed on the manipulation type distribution, which may vary across branches due to random sampling. Subsequently, the heads are concatenated as described in Subsection 2.4.2.4, and only the newly added network components are trained while all other weights remain fixed. Parameter initialization for the final fully connected layers follows the default Keras settings, and no seed is specified for dropout layers. In the final stage, the entire neural network is trained using a very low learning rate of 0.00001.

In order to identify the optimal model, a grid search over a range of hyperparameters is typically necessary. However, for the purposes of this study, we present the results obtained using *Adam* as optimizer with a learning rate of 0.001 for FF++ and 0.0002 for CelebDF (v2), as these values yielded satisfactory performance in preliminary experiments. We applied a batch size of 32 and the *Reduce on Plateau* scheduler: we check the validation set loss each epoch and reduce the learning rate by a factor of 0.1 when it does not improve. To avoid long training sessions, we also applied an early stopping method: during training, the model is overwritten whenever it achieves a lower loss on the validation set, but if the loss does not decrease for 3 consecutive epochs, training stops. All experiments were conducted on an Ubuntu 22.04.2 LTS using Keras and Tensorflow in a Python environment. The GPU used is an NVIDIA RTX A6000 50 GB.

2.4.2.8 Results

A comprehensive set of experiments was conducted to evaluate the performance of the proposed approach under varying compression conditions and architectural configurations. In accordance with established protocols in deepfake detection, frame-level accuracy (ACC) and area under the receiver operating characteristic curve (AUC) are reported. Video-level accuracy is determined by applying majority voting to all frame-level predictions within each video clip.

To minimize variability resulting from random initialization and training dynamics, all results are averaged over five runs using fixed training data. In addition to reporting mean values and standard deviations, the statistical significance of differences between model variants is assessed using the Wilcoxon signed-rank test. For each pairwise comparison, the p-value and effect size, measured by Cohen's d, are reported. All statistical comparisons are conducted across multiple seeds to account for randomness in the training process.

2.4.2.8.1 Impact of unseen compressed data on model performance

As reported above, detecting deepfakes in compressed environments is a well-documented challenge. Existing research shows that deepfake detection model performance significantly degrades under lossy compression [149]. To establish a clear starting point, baseline models were used to quantify the impact of compression within the experimental setup. This approach underscores the severity of the issue and provides a foundation for demonstrating the improvements achieved by the proposed framework in subsequent experiments. Specifically, we trained all the selected baseline models on frames extracted from videos with three different compression levels (RAW, HQ, and LQ), evaluating their performance across all these levels. However, to provide a focused illustration of the results, we present the detailed findings for MobileNetV3Small in Table 2.7, though they are representative of the trends observed across the other baseline architectures considered in our analysis. The results reveal a significant drop in accuracy when the model is tested on compression levels not seen during training. In particular, models trained on RAW frames, from FF++ and Celeb-DF (v2) datasets, exhibited the most substantial decrease in performance when evaluated on LQ frames, as evidenced by a drop in video accuracy from 99.86% to 34.57% on the FF++ dataset, with similar trends observed on Celeb-DF (v2). Models trained on HQ frames lost accuracy when tested on LQ frames, but performed reasonably well on RAW frames.

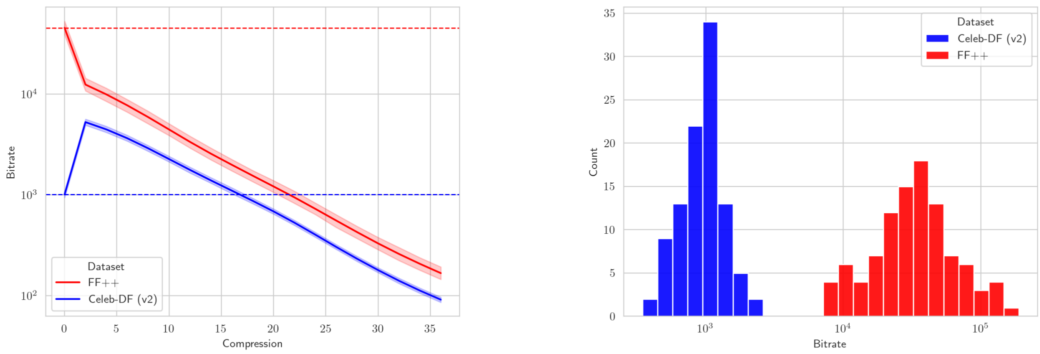
Table 2.7: *Performance of the MobileNetV3Small model trained on different compression levels (LQ, HQ, and RAW) and evaluated on FF++ (a) and Celeb-DF (v2) (b) datasets. Best results are highlighted in bold.*

(a)

FF++ test Set									
Training Set	LQ			HQ			RAW		
	FRAME ACC	AUC	VIDEO ACC	FRAME ACC	AUC	VIDEO ACC	FRAME ACC	AUC	VIDEO ACC
LQ	**82.68 ± 1.29**	**84.97 ± 0.37**	**86.11 ± 0.41**	82.33 ± 1.09	85.71 ± 1.09	85.01 ± 2.72	80.34 ± 3.18	83.41 ± 3.18	83.09 ± 3.24
HQ	63.90 ± 1.66	78.22 ± 0.33	66.76 ± 3.24	**94.33 ± 0.04**	**98.25 ± 0.04**	**96.85 ± 0.12**	96.32 ± 0.55	99.04 ± 0.55	97.33 ± 0.65
RAW	37.28 ± 1.56	57.95 ± 0.54	34.57 ± 1.26	39.22 ± 1.39	82.36 ± 1.39	37.92 ± 1.20	**99.50 ± 0.04**	**99.98 ± 0.04**	**99.86 ± 0.00**

(b)

Celeb-DF (v2) test set									
Training Set	LQ			HQ			RAW		
	FRAME ACC	AUC	VIDEO ACC	FRAME ACC	AUC	VIDEO ACC	FRAME ACC	AUC	VIDEO ACC
LQ	**83.08 ± 0.35**	**91.00 ± 0.77**	**86.42 ± 1.46**	87.94 ± 0.75	95.29 ± 0.32	90.99 ± 0.95	88.22 ± 0.83	95.50 ± 0.32	91.18 ± 0.96
HQ	63.23 ± 1.94	84.86 ± 0.57	62.48 ± 2.67	**95.60 ± 0.68**	**98.91 ± 0.18**	98.07 ± 0.42	96.18 ± 0.67	99.11 ± 0.15	98.52 ± 0.24
RAW	53.81 ± 3.23	79.65 ± 1.00	51.16 ± 4.38	95.15 ± 0.50	98.85 ± 0.25	**98.13 ± 0.55**	**96.64 ± 0.60**	**99.39 ± 0.13**	**98.65 ± 0.54**



(a) **(b)**

Figure 2.15: *(a) Bitrate as a function of CRF for 100 randomly selected real videos from the FF++ and Celeb-DF (v2) datasets. The solid lines represent the average bitrate, while the shaded areas indicate the 95% confidence interval across the selected videos. (b) Distribution of the bitrate for 100 raw videos from the Celeb-DF (v2) and FF++ datasets, highlighting the disparity in their initial quality levels.*

Training with LQ frames demonstrated the most stability across various compression levels, yet made it difficult to achieve competitive performance on higher-quality frames. However, the results obtained on the Celeb-DF (v2) dataset present a different behaviour. Although we replicated the same compression levels (LQ, HQ, and RAW) as in FF++, the performance drop from RAW to HQ observed before is not evident in this case. Specifically, the model trained on RAW frames did not show a significant drop in accuracy when tested on HQ frames, suggesting that the original videos in the Celeb-DF (v2) dataset may have already undergone some degree of compression before we applied our artificial levels. To corroborate this observation, we analyzed the video bitrate as a function of the CRF for both datasets, as shown in Fig. 2.15a. The bitrate, which measures the amount of data encoded per second of video, provides insight into the compression level applied; lower bitrates generally indicate higher compression levels. The FF++ data exhibits a typical pattern, with the video bitrate decreasing steadily as the CRF increases, reflecting the effects of increasing compression. In contrast, the starting quality of Celeb-DF (v2) aligns more closely with a CRF value between 15 and 20, which is much nearer to the HQ setting ($c = 23$) rather than to the RAW setting ($c = 0$). Consequently, the curve initially rises as CRF increases from 0, reflecting an increase in bitrate due to the minimal compression applied before aligning with the expected compression behavior in the later stages of the curve. Fig. 2.15b further highlights the lower starting quality of Celeb-DF (v2) with a clear distinction in the raw video bitrate distributions between the two datasets.

Table 2.8: *Comparison between vc and fc augmentation strategies on FF++ dataset. The values in the columns represent the mean on the LQ, HQ, and RAW test sets. Best results are highlighted in bold.*

MODEL	FRAME ACC.	AUC	VIDEO ACC.
MobileNetV3Small (vc)	87.19	92.61	89.91
MobileNetV3Small (fc)	**88.50**	**93.45**	**91.18**
MobileNetV3Large (vc)	**90.74**	94.48	**93.43**
MobileNetV3Large (fc)	90.67	**94.94**	93.23
EfficientNetV2B0 (vc)	88.49	93.88	92.11
EfficientNetV2B0 (fc)	**89.39**	**94.56**	**92.29**
Xception (vc)	89.21	94.08	91.87
Xception (fc)	**90.39**	**95.14**	**93.42**

2.4.2.8.2 Data augmentation strategies comparison

Compression-aware augmentation constitutes a central element of the proposed framework. Two sampling strategies were defined to assign compression levels during training: the *vc* and the *fc* approaches. To evaluate the impact of these strategies, a comparative analysis was performed across the four backbone architectures utilized in this study. Two versions of each model were trained on the FF++ dataset under identical conditions, differing only in the sampling strategy used to generate the augmented training set. The results are presented in Table 2.8. A consistent trend emerges in favor of the *fc* approach concerning AUC, indicating a more effective encoding of compression variability during training. Despite minor variations in frame and video-level accuracy, the improvement is observed across all architectures, suggesting that exposure to a broader and more uniform compression spectrum results in more stable performance across decision thresholds. Based on this empirical evidence, the *fc* configuration is adopted as the standard setting in all subsequent experiments.

2.4.2.8.3 Evaluation of the Multi-Head Network architecture

Following the analysis of baseline model limitations in Subsection 2.4.2.8.1, this section evaluates the impact of the proposed augmentation technique and the Multi-Head Network architecture on the robustness and accuracy of deepfake detection models. Experimental results are summarized in Tables 2.9a-2.9b. Compared to models trained on isolated compression levels (see Table 2.7), models trained with the compression-aware augmentation, labeled as *Simple* in the tables, show superior adaptability across quality settings. Unlike models specialized on LQ, HQ, or RAW videos, which overfit and perform poorly outside their domain, the *Simple* models benefit from exposure to a broader compression spectrum, resulting in higher accuracy and more balanced performance.

Table 2.9: *Performance comparison between standard models and their enhanced versions with our MHN across different compression levels (LQ, HQ, and RAW) on the FF++ (a) and Celeb-DF (v2) (b) datasets. Best results are highlighted in bold.*

(a)

MODELS	**Million Params**	**FaceForensics++ test set**								
		LQ			**HQ**			**RAW**		
		FRAME ACC	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**
MobileNetV3Small - Simple	0.94	80.50 ± 2.38	86.70 ± 1.97	84.50 ± 2.80	90.27 ± 1.99	95.31 ± 1.22	92.98 ± 2.03	94.74 ± 1.43	98.35 ± 0.76	96.07 ± 1.11
MobileNetV3Small - MH	0.98	**83.02 ± 0.66**	**88.74 ± 0.36**	**87.68 ± 1.16**	**92.63 ± 0.30**	**97.27 ± 0.12**	**95.44 ± 0.59**	**98.25 ± 0.13**	**99.81 ± 0.03**	**98.62 ± 0.41**
MobileNetV3Large - Simple	3.00	81.89 ± 2.95	87.87 ± 1.57	87.08 ± 2.44	92.48 ± 1.17	96.85 ± 0.89	95.05 ± 1.07	96.27 ± 1.03	99.15 ± 0.39	97.19 ± 0.93
MobileNetV3Large - MH	3.12	**84.44 ± 0.85**	**90.68 ± 1.38**	**90.03 ± 1.06**	**94.39 ± 0.93**	**98.33 ± 0.48**	**96.36 ± 0.79**	**98.90 ± 0.21**	**99.91 ± 0.04**	**99.11 ± 0.19**
EfficientNetV2B0 - Simple	5.92	76.98 ± 2.35	86.62 ± **0.24**	81.55 ± 2.96	91.41 ± 0.91	97.06 ± **0.12**	94.38 ± 1.20	96.81 ± **0.21**	99.33 ± 0.08	97.62 ± 0.31
EfficientNetV2B0 - MH	6.08	**83.50 ± 1.97**	**89.81** ± 0.85	**89.91 ± 1.66**	**94.79 ± 0.71**	**98.33** ± 0.35	**97.02 ± 0.51**	**98.99** ± 0.24	**99.93 ± 0.03**	**99.23 ± 0.17**
Xception - Simple	20.86	80.82 ± 2.89	88.08 ± 1.11	86.71 ± 3.28	93.13 ± 1.95	97.79 ± 0.37	95.64 ± 1.55	97.22 ± **0.24**	99.55 ± 0.06	97.91 ± **0.11**
Xception - MH	21.12	**84.23 ± 2.14**	**90.48 ± 0.97**	**89.11 ± 1.96**	**94.66 ± 0.70**	**98.52 ± 0.27**	**96.45 ± 0.58**	**99.45** ± 0.31	**99.97 ± 0.02**	**99.71** ± 0.31

(b)

MODELS	**Million Params**	**Celeb-DF (v2) test set**								
		LQ			**HQ**			**RAW**		
		FRAME ACC	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**
MobileNetV3Small - Simple	0.94	82.55 ± 4.38	91.70 ± 0.81	86.14 ± 5.27	91.79 ± 1.12	98.09 ± 0.23	94.02 ± 1.71	92.70 ± 0.91	98.41 ± 0.24	94.90 ± 1.34
MobileNetV3Small - MH	0.98	**86.64 ± 0.66**	**93.77 ± 0.53**	**90.46 ± 0.23**	**94.92 ± 0.40**	**98.94 ± 0.15**	**96.80 ± 0.60**	**96.31 ± 0.33**	**99.40 ± 0.08**	**98.03 ± 0.28**
MobileNetV3Large - Simple	3.00	83.62 ± 3.39	92.25 ± 1.30	89.04 ± 3.61	94.23 ± 0.99	98.62 ± **0.26**	96.64 ± 0.58	94.89 ± 0.96	98.93 ± 0.20	97.30 ± 0.91
MobileNetV3Large - MH	3.12	**87.48 ± 1.66**	**94.51 ± 0.92**	**91.82 ± 1.03**	**96.16 ± 0.68**	**99.11** ± 0.30	**97.91 ± 0.37**	**97.16 ± 0.80**	**99.63 ± 0.09**	**98.49 ± 0.86**
EfficientNetV2B0 - Simple	5.92	**85.03** ± 1.03	91.97 ± 0.42	89.50 ± 0.50	94.13 ± **0.30**	98.70 ± 0.18	96.53 ± 0.36	94.75 ± **0.30**	98.96 ± 0.17	96.95 ± 0.33
EfficientNetV2B0 - MH	6.08	84.78 ± **0.58**	**92.12 ± 0.16**	**90.15 ± 0.44**	**95.44** ± 0.34	**99.16 ± 0.09**	**97.95 ± 0.34**	**96.36** ± 0.41	**99.52 ± 0.04**	**98.22 ± 0.31**
Xception - Simple	20.86	85.37 ± 1.66	93.30 ± 0.55	90.69 ± 2.06	95.22 ± 0.62	**99.30 ± 0.14**	97.11 ± 0.66	96.00 ± **0.46**	99.47 ± **0.10**	97.72 ± 0.56
Xception - MH	21.12	**86.44 ± 0.83**	**93.68 ± 0.39**	**91.16 ± 0.83**	**95.81 ± 0.43**	99.24 ± 0.25	**97.80 ± 0.40**	**96.90** ± 0.47	**99.60** ± 0.12	**98.38 ± 0.38**

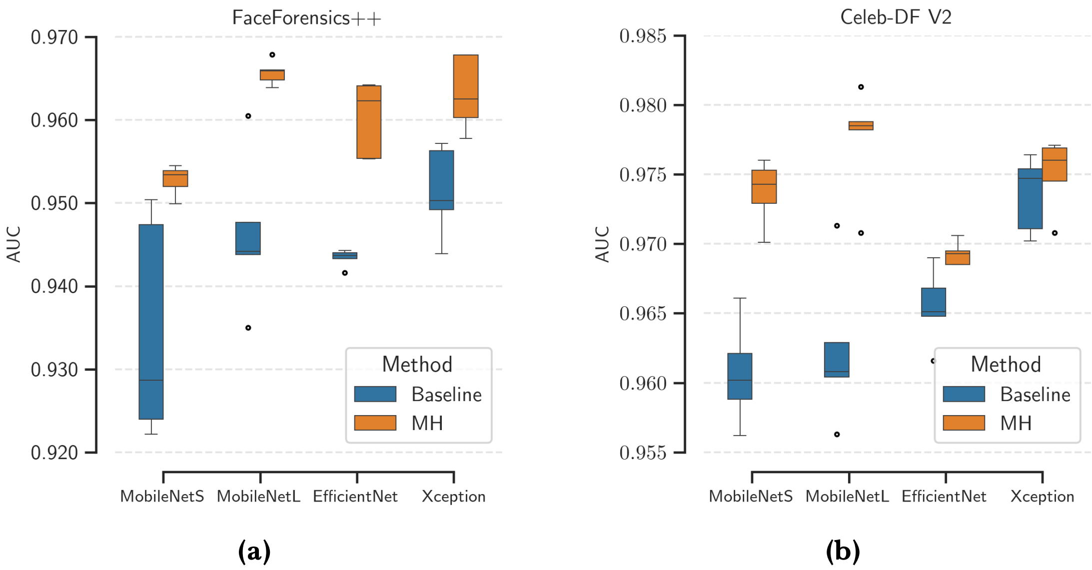
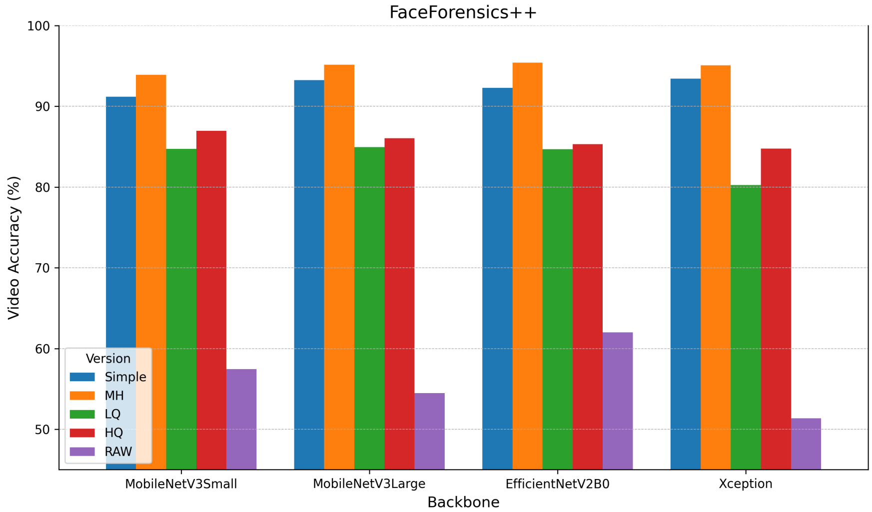


Figure 2.16: *Boxplots of AUC scores for baseline (Simple) and MH-enhanced models, computed over five training runs on the FF++ dataset (a) and the Celeb-DF (v2) dataset (b).*

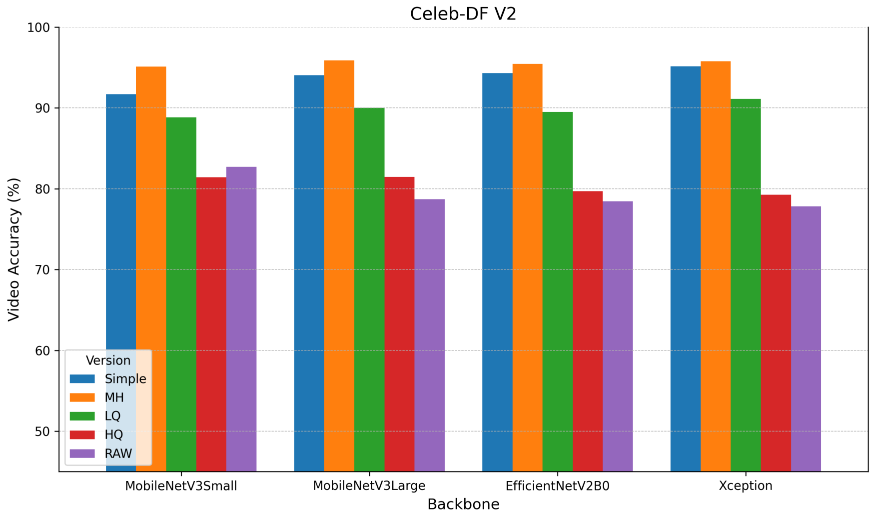
This robustness is evident across both datasets, although the accuracy metrics for Celeb-DF (v2) are slightly skewed due to its lower quality. As previously discussed, the training labels were based on FF++ compression levels, but Celeb-DF (v2) videos were likely of lower quality, which may have affected the results. Uniform labeling was maintained across datasets to ensure consistency in evaluation.

These results indicate that the augmentation technique alone is insufficient. A specialized architecture, such as the MHN, is required to leverage the injected compression information fully. Incorporating the MHN yields a significant performance improvement over the *Simple* model that uses only augmentation. Additionally, the MHN stabilizes the training process, as evidenced by lower standard deviations across multiple training sessions and more consistent results (Fig. 2.16). Figure 2.17 presents the mean video accuracy across the three test sets for each configuration. MHN-enhanced models consistently outperform both *Simple* models and single-compression-trained baselines, confirming the effectiveness of combining compression-aware augmentation with a dedicated architectural design.

A representative case is MobileNetV3Small with MHN, which outperforms the Simple configurations of its larger baseline, MobileNetV3Large, and nearly matches the performance of more complex, parameter-heavy models like EfficientNet and Xception. Such findings suggest that smaller models can challenge the conventional approach of increasing model size to achieve better performance when strategically enhanced. The advantage of MHN is also evident across compression settings.



(a)



(b)

Figure 2.17: *Comparison of the mean video accuracy, averaged across all three compression levels (LQ, HQ, and RAW), for the FF++ (a) and Celeb-DF (v2) (b) datasets, using different model architectures and five training approaches: Simple (augmentation only), MH (Multi-Head architecture), and models trained on RAW, HQ, or LQ images.*

As discussed in Section 2.4.2.8.1, models trained exclusively on RAW suffer substantial degradation when evaluated on compressed data, while those trained on HQ or LQ generalize more effectively, especially on Celeb-DF (v2), where native compression artifacts may act as a form of implicit regularization.

To assess the robustness and significance of the observed improvements, we performed a statistical comparison between each baseline model and its MHN-

Table 2.10: *Effect sizes (Cohen's d) for AUC improvements between MHN-enhanced models and their baselines on FF++ and CelebV2.*

Dataset	Model	Cohen's d
FF++	MobileNetV3Small	1.321
	MobileNetV3Large	2.201
	EfficientNetV2B0	**3.456**
	Xception	1.330
CelebV2	MobileNetV3Small	**2.183**
	MobileNetV3Large	2.011
	EfficientNetV2B0	1.599
	Xception	1.131

enhanced counterpart. As shown in Table 2.10, all pairwise comparisons yielded statistically significant AUC differences ($p = 0.031$), below the commonly adopted $\alpha = 0.05$ threshold used to reject the null hypothesis of no difference [10]. Beyond significance, the strength of the observed improvements was quantified using Cohen's d, a standardized measure of effect size. The magnitude of the observed improvements, expressed through Cohen's d, consistently exceeds 1.1, indicating large to very large effects according to standard interpretation criteria [29]. The most substantial gains are observed for EfficientNetV2B0 on FF++ ($d = 3.46$) and MobileNetV3Small on CelebV2 ($d = 2.18$), showing that MHN yields consistent and practically meaningful improvements across datasets and architectures.

The next section investigates whether architectural gains can be further amplified by exploiting structured parallelism, as opposed to increasing model depth or width.

2.4.2.8.4 Evaluation of the Multi-Branch Network architecture

Building on the concept of efficient parameter utilization, this section investigates whether the Multi-Branch Network architecture can serve as a more effective alternative to increasing model depth by leveraging the parallelism of multiple network branches. This analysis was conducted on the FF++ dataset, selecting the MobileNetV3 family due to its balance between performance and efficiency.

As shown in Table 2.11, the adoption of additional branches leads to progressive improvements across all compression conditions, particularly in terms of AUC (Fig. 2.18). Notably, the transition from a single-branch (MH) to a two-branch architecture (MB2) yields measurable gains with minimal increase in model complexity, consistently improving AUC even under aggressive compression. While subsequent branches continue to enhance performance, the improvements saturate quickly and become less proportional to the added computational cost.

Table 2.11: *Performance comparison between standard models and their enhanced versions with our MHN and MBN across different compression levels (LQ, HQ, and RAW), for MobileNetV3Small (a) and MobileNetV3Large (b) on the FF++ dataset.*

(a)

Models	**Million Params**	**MobileNet V3 Small - Test Set (FF++ Dataset)**								
		LQ			**HQ**			**RAW**		
		FRAME ACC	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**
Simple	0.94	80.50 ± 2.38	86.70 ± 1.97	84.50 ± 2.80	90.27 ± 1.99	95.31 ± 1.22	92.98 ± 2.03	94.74 ± 1.43	98.35 ± 0.76	96.07 ± 1.11
MH	0.98	83.02 ± 0.66	88.74 ± 0.36	87.68 ± 1.16	92.63 ± 0.30	97.27 ± **0.12**	95.44 ± 0.59	98.25 ± 0.13	99.81 ± 0.03	98.62 ± 0.41
MH MB2	1.95	84.74 ± **0.22**	89.66 ± 0.33	88.85 ± 0.65	93.51 ± **0.13**	97.51 ± 0.20	96.22 ± 0.25	98.43 ± **0.12**	99.84 ± 0.05	98.91 ± **0.21**
MH MB3	2.93	85.00 ± 0.28	89.90 ± 0.13	89.37 ± 0.70	93.70 ± 0.15	97.61 ± 0.19	**96.30** ± 0.17	98.54 ± 0.19	99.86 ± 0.04	98.85 ± 0.33
MH MB4	3.91	85.31 ± 0.39	90.02 ± 0.12	**89.40** ± **0.22**	93.80 ± 0.17	97.68 ± 0.21	96.22 ± **0.15**	98.62 ± 0.19	99.88 ± **0.02**	99.05 ± 0.33
MH MB5	4.88	**85.47** ± 0.35	**90.24** ± **0.08**	89.17 ± 0.27	**93.84** ± 0.14	**97.72** ± **0.12**	96.28 ± 0.20	**98.65** ± 0.19	**99.89** ± 0.03	**99.08** ± 0.39

(b)

Models	**Million Params**	**MobileNet V3 Large - Test Set (FF++ Dataset)**								
		LQ			**HQ**			**RAW**		
		FRAME ACC	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**	**FRAME ACC**	**AUC**	**VIDEO ACC**
Simple	3.00	81.89 ± 2.95	87.87 ± 1.57	87.08 ± 2.44	92.48 ± 1.17	96.85 ± 0.89	95.05 ± 1.07	96.27 ± 1.03	99.15 ± 0.39	97.19 ± 0.93
MH	3.12	84.44 ± 0.85	90.68 ± 1.38	90.03 ± 1.06	94.39 ± 0.93	98.33 ± 0.48	96.36 ± 0.79	98.90 ± 0.21	99.91 ± 0.04	99.10 ± 0.19
MH MB2	6.24	86.74 ± 0.54	91.64 ± 0.11	91.14 ± 0.64	95.47 ± **0.09**	98.57 ± 0.12	96.87 ± 0.17	99.15 ± 0.11	99.96 ± **0.01**	99.40 ± 0.17
MH MB3	9.36	86.92 ± 0.49	91.85 ± **0.09**	91.23 ± 0.78	95.48 ± 0.13	98.64 ± 0.09	96.85 ± 0.13	99.25 ± 0.11	99.96 ± **0.01**	99.43 ± 0.13
MH MB4	12.48	87.19 ± **0.19**	92.04 ± **0.09**	91.26 ± 0.66	95.56 ± 0.12	98.72 ± 0.11	**96.96** ± **0.10**	99.27 ± 0.09	99.96 ± **0.01**	99.40 ± 0.17
MH MB5	15.60	**87.47** ± 0.20	**92.18** ± 0.10	**91.52** ± **0.59**	**95.65** ± 0.10	**98.76** ± **0.08**	96.88 ± 0.11	**99.36** ± **0.07**	**99.97** ± **0.01**	**99.57** ± **0.09**

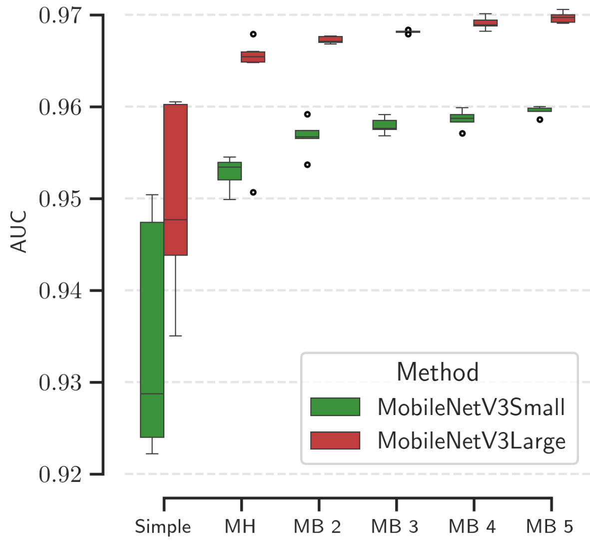


Figure 2.18: *Boxplots of AUC scores for baseline (Simple), MHN, and MBN, evaluated on the FF++ dataset using MobileNetV3 family as backbone. Results are averaged over five independent training runs.*

Table 2.12: *Effect sizes (Cohen's d) computed on the AUC values obtained from the FF++ dataset. Each MBN variant is compared against the corresponding Simple model from the MobileNetV3 family.*

Model	**Variant**	**Cohen's d**
Small	MB2	1.88
	MB3	1.87
	MB4	1.93
	MB5	1.93
Large	MB2	1.68
	MB3	1.73
	MB4	1.91
	MB5	1.95

Nevertheless, the ensemble effect introduced through bootstrap resampling also contributes to lower variance across training runs, suggesting more stable optimization dynamics. Moreover, the MBN's modular design allows practitioners to fine-tune the trade-off between accuracy and complexity, making it adaptable to diverse deployment scenarios. Also in this case, we conducted a statistical analysis comparing each MBN variant to the baseline Simple model: all configurations show statistically significant differences (Wilcoxon signed-rank test, $p = 0.0312$), with large effect sizes across the board (Cohen's d ranging from 1.68 to 1.95, see Table 2.12).

Table 2.13: *Performance comparison of various state-of-the-art deepfake detection approaches on the FF++ dataset, evaluated on LQ and HQ test sets. The table reports the area under the ROC curve (AUC %). Each approach is listed with its respective backbone architecture in parentheses.*

Method	**AUC (%)**	
	LQ	**HQ**
Xception [149]	89.30	96.30
F^3-Net [136] (Xception)	93.30	98.10
SPSL [104] (Xception)	82.82	95.32
FDFL [98] (Xception)	92.40	99.30
MAT [197] (EfficientNet-B4)	90.40	99.29
SIA [164] (EfficientNet-B4)	93.45	99.35
RECCE [16] (Xception)	95.02	99.32
BOF [119](Vision Transformer [43])	91.61	99.36
LRL [23] (Custom)	95.21	99.46
MLR [185] (MC3-18 [175])	96.18	98.27
Our (MobileNetV3Large - MH MB5)	92.18	98.76

2.4.2.8.5 Comparison with State-Of-The-Art methods

To evaluate the competitiveness of our approach, we selected the MobileNetV3-Large with five branches (MH MB5), identified as the best performer in the previous section, for comparison against state-of-the-art (SOTA) deepfake detection methods. The evaluation was conducted on the FF++ dataset under both LQ and HQ compression conditions (Table 2.13).

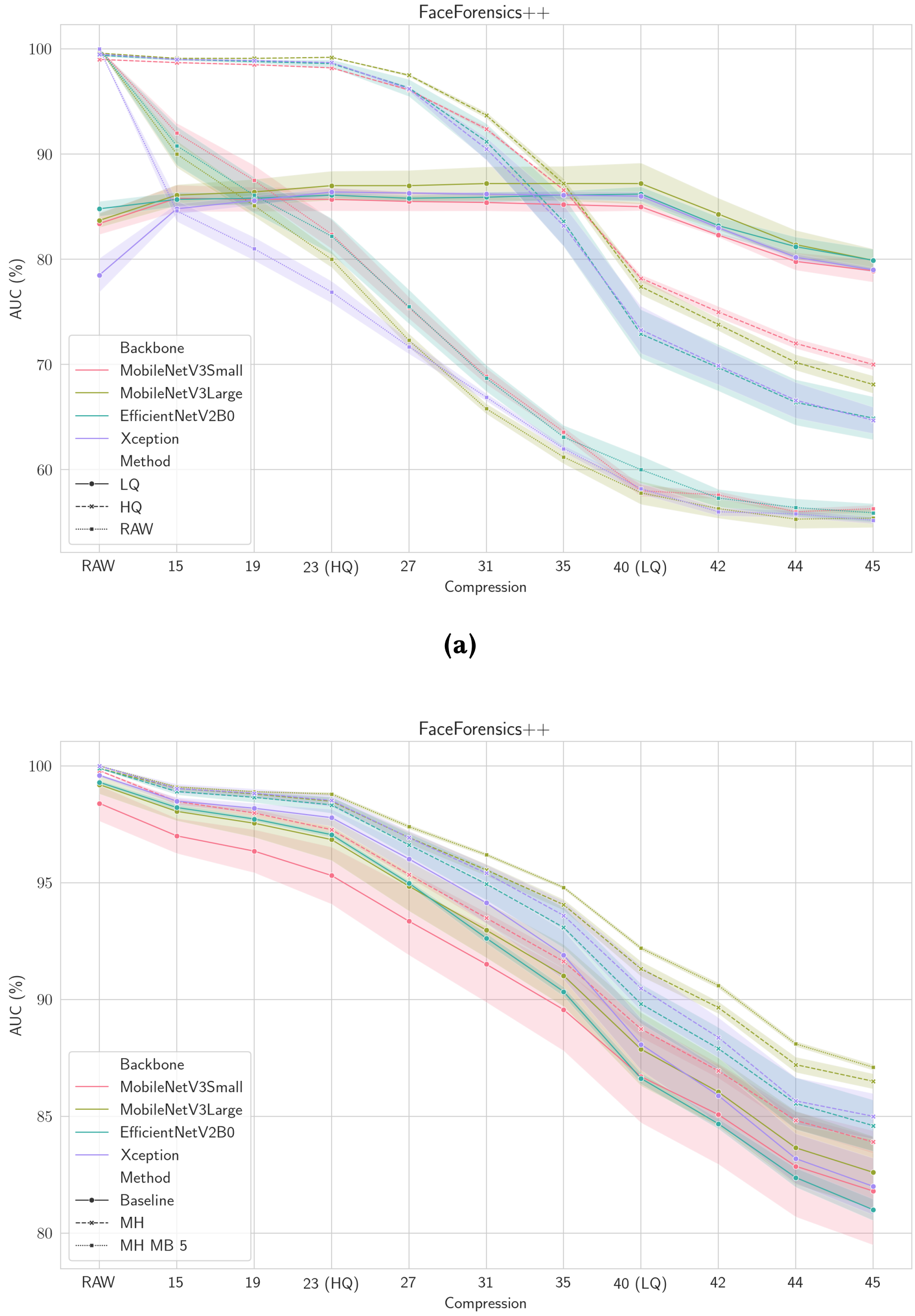
While the objective is not to outperform SOTA methods, our results show that a lightweight architecture such as MobileNetV3, when enhanced with our Multi-Head and Multi-Branch framework, can achieve comparable performance to more complex and computationally demanding models like Xception and EfficientNet-B4. Notably, our architecture reaches an AUC of 92.18% on LQ and 98.76% on HQ, values that are in line with those reported by recent top-performing methods.

While most existing approaches evaluate detection models under only two fixed compression settings (LQ and HQ from the FF++ benchmark), our framework is designed to be robust beyond such discrete configurations. Although we report results on these standard levels for comparison purposes, our method is rooted in a more comprehensive and realistic modeling of compression variability. As illustrated in Fig. 2.19a, disregarding the underlying compression distribution may lead to substantial generalization gaps. For instance, models trained solely on low-quality (LQ) compressed data tend to exhibit improved robustness across a wider range of compression levels, albeit with lower peak performance. Conversely, models trained on high-quality (HQ) compressed data often achieve high accuracy on similar input distributions but suffer from significant performance degradation when exposed to stronger compression artifacts. In this context, the performance achieved by our method does not result from explicit optimization for these specific compression levels (Fig. 2.19b). Rather, it reflects the intrinsic robustness of the proposed framework, emerging from the combination of a dynamic MPEG-based augmentation strategy with a modular architectural design.

Finally, it is worth emphasizing that the proposed framework is inherently backbone-agnostic. Its components can be readily integrated into a wide range of architectures and deployment scenarios, offering a general-purpose enhancement strategy for deepfake detection pipelines. This adaptability makes it a promising foundation for future research aimed at maintaining high detection performance under realistic, resource-constrained operating conditions.

2.4.2.9 Conclusions

In this section, we presented an innovative and modular approach to deepfake detection that explicitly addresses the degradation effects introduced by video compression, a major challenge in real-world applications. Our method integrates a Multi-Head Network and Multi-Branch Network architecture with a compression-aware data augmentation strategy, designed to improve performance across a range of compression levels. Experimental results on the FF++ and Celeb-DF (v2) datasets show that the MHN architecture yields consistent gains in detection accuracy, particularly under low-quality conditions. Furthermore, the MBN architecture demonstrates that performance can be improved by adding parallel branches, providing a scalable alternative to depth-based expansion and enabling a more flexible balance between accuracy and model complexity. The proposed components are compatible with a wide range of network designs and can be integrated into existing architectures with minimal structural modifications. This makes the approach a practical candidate for deployment in scenarios with constrained computational resources or diverse compression settings, providing a unified and effective solution to the problem of detecting compressed deepfakes.



(a)

(b)

Figure 2.19: *AUC scores across varying compression levels on the FF++ dataset. Each line represents a different training setup: (a) models trained on a single compression level (RAW, HQ, LQ); (b) models trained using the proposed fc strategy. Shaded areas denote the standard deviation computed over five training runs.*

Chapter 3

Audio Adversarial Attacks

In the previous chapter, we examined how compression weakens the performance of deepfake detectors. This issue is particularly concerning because compression is ubiquitous in multimedia sharing, as most applications reduce file sizes for faster transmission. Unlike this unintentional degradation, the present chapter investigates vulnerabilities in machine learning systems under deliberate adversarial attacks. Focusing on Voice Disorder Detection system vulnerabilities, this chapter introduces the topic of audio adversarial attacks without delving into implementation details. Instead, it provides an overview of how these attack techniques pose a threat to machine learning models. Starting with speech recognition systems, and moving on to voice recognition systems, we review a series of studies related to the threat of audio adversarial attacks, culminating in our publication [131], which addresses this issue for the first time in voice disorder detection systems.

3.1 Introduction

The recent development of AI systems has significantly accelerated the growth of speech and voice recognition technologies. These systems are now integrated into a wide range of everyday applications, from digital voice assistants such as Siri, Alexa, or Google Assistant, to voice-based biometric authentication, and even automated speech quality and health assessment platforms. Their widespread adoption stems from their ability to enable natural human–machine interactions, reducing the need for textual or physical input and allowing more immediate and inclusive access to the functionalities of intelligent devices [138, 190].

However, this growing adoption has also highlighted a series of emerging risks. Modern speech recognition and audio classification systems rely on ML and deep learning models, which, while achieving impressive performance under standard conditions, can be vulnerable to intentionally crafted perturbations designed to mislead them [19, 166]. These perturbations, known as adversarial attacks, consist of minimal and often imperceptible modifications to the original audio signal that induce the model to produce incorrect classifications or behave unpredictably.

In the audio domain, adversarial attacks can take several forms. They may aim to spoof another user's voice (thus bypassing a voice authentication system), cause a voice assistant to execute unauthorized commands, or, in healthcare, manipulate the output of automatic systems that analyze pathological or disordered speech [1]. Unlike the visual domain [84], where perturbations are localized at the pixel level, audio perturbations must satisfy strict perceptual constraints, since even minor changes in the waveform or spectrogram can affect the quality or intelligibility of the sound. This makes audio adversarial design particularly challenging, and, paradoxically, more dangerous, since well-crafted perturbations can remain entirely imperceptible to human listeners [90, 106].

Over the past few years, the literature has identified several categories of adversarial attacks in the audio domain. Some operate directly in the time domain, modifying the raw waveform, while others act in the frequency domain, manipulating representations such as the spectrogram or Mel-Frequency Cepstral Coefficients (MFCC). Adversarial attacks can also be broadly categorized based on the attacker's knowledge of the target model. In a white-box attack scenario, the attacker possesses complete knowledge of the model's architecture and parameters, enabling precise and potent adversarial example crafting [166]. In this scenario, gradient-based methods, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), leverage the model's gradient information to create perturbations that maximize prediction error while remaining imperceptible to human ears [110]. In contrast, black-box attacks assume that the attacker has no direct access to the target model information but can observe its output given specific inputs, making these attacks more reflective of real-world scenarios [126]. Several works exploit this strategy to fool audio-based systems [56, 199, 63, 137]. Moreover, beyond digital attacks, researchers have begun exploring physical-world scenarios, where adversarial audio is played through speakers in real environments to assess model robustness under realistic acoustic conditions [183].

The implications of these vulnerabilities are profound, particularly in contexts where voice carries sensitive or clinical information. One emblematic example is voice disorder detection systems, which employ machine learning algorithms to identify vocal pathologies or phonatory disorders [64, 189]. The possibility that an adversarial attack could alter or falsify an automatic diagnosis raises significant ethical, security, and reliability concerns, especially given the increasing adoption of telemedicine and remote speech analysis solutions [40].

Understanding the characteristics and vulnerabilities of speech recognition models is, therefore, a fundamental step in ensuring the security and reliability of voice-based applications. For voice disorder detection systems, the challenge is twofold: ensuring robust and accurate recognition of vocal disorders under conditions of noise and individual variability, and guaranteeing that models cannot be manipulated by intentional attacks that compromise their diagnostic reliability [193].

The following section provides an overview of voice disorder pathology and the tools that AI makes available in medical applications. Then, the state of the art in voice disorder detection systems will be presented, analyzing their evolution and common architectures. It will then introduce our original contribution, focused on the analysis of voice disorder detection vulnerabilities.

3.2 Voice Disorder

Voice disorders manifest as variations from normal voice quality, pitch, and loudness concerning an individual's age, gender, and cultural background [13]. These disorders impact a substantial portion of the population, with estimates suggesting that up to 20% of people may encounter a voice disorder during their lifetime [73]. Severity ranges from mild inconvenience to complete voice loss, and the underlying causes may be functional or, in some cases, cancerous. Diagnosis generally requires a comprehensive clinical assessment, encompassing interviews, listening tests, acoustic analyses, and laryngoscopy. When cancer is suspected, a biopsy is carried out [176]. Given their prevalence and diversity, diagnosing voice disorders can be time-consuming for both patients and healthcare professionals. This lengthy process often leads to higher costs for healthcare systems [30]. Delays in diagnosis may worsen the disorder and complicate subsequent treatment [128].

Integrating machine learning into healthcare has driven major advances in computer-aided diagnosis systems, supporting clinicians with computer-generated insights [5]. These models are now widely used to analyze imaging data, including MRI and CT scans, for detecting cancers, fractures, and various other medical issues. In addition, they enable real-time patient monitoring and help anticipate critical events like sepsis or cardiac failure [7]. Within this wide array of uses, automated voice disorder detectors have demonstrated considerable promise. By extracting and analyzing acoustic features from recordings, these systems distinguish between healthy and pathological voices [66]. However, the integration of such technology introduces new vulnerabilities. Particularly concerning is the potential for adversarial manipulation, a concept well-documented in other ML domains, such as image and speech recognition [19].

Adversarial attacks involve deliberately altering input data to deceive the model into erroneous predictions [182]. These manipulations are often imperceptible or seemingly benign to human listeners, yet they can drastically alter the model's output. An especially concerning application of such attacks could be the creation of manipulated audio recordings designed to falsely suggest the presence of a medical condition in a target individual. Such deceptive practices could have profound implications: a person might be unjustly disqualified from employment opportunities or be subjected to inflated health insurance rates based on fabricated evidence of illness, and a public figure could suffer damage to their reputation. For instance, fake pathological recordings could be used in cyberbullying or harassment campaigns to manipulate public perception of individuals [194].

In the following section, we explore adversarial attacks and audio manipulations within the context of Voice Disorder Detection (VDD). By identifying and analyzing the potential threats posed by adversarial or manual manipulations, we aim to highlight the critical need for robust countermeasures. Ensuring the security and integrity of VDD systems against manipulations is paramount in safeguarding diagnostic accuracy and trust in these tools. Section 3.4.1 describes the attacks implemented to assess the robustness of voice disorder detection systems. Section 3.4.1.4 describes the protocol used to conduct our evaluation, while section 3.4.1.5 reports the obtained results. Finally, conclusions are drawn in Section 3.4.1.7.

3.3 Voice Disorder Detections

In recent years, the field of voice disorder detection has made significant progress, with the development of various methodologies aimed at more accurately distinguishing between healthy and pathological voices. Central to these advancements is applying ML algorithms to vocal data analysis. These algorithms utilize a range of classifiers and leverage key acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCC) [48] or Multidimensional Voice Program (MDVP) [4] to enhance diagnostic precision. Among the ML techniques, SVM has seen widespread use due to its effectiveness in classifying pathological and healthy voices. Studies employing SVM with MFCC have reported significant accuracies, though often limited by small datasets [75]. For instance, Al-Dhief et al. [3] utilized SVM with MFCC methodology, achieving an 91.17% accuracy using voice signals from the Saarbrücken voice database (SVD)[1]. Souissi et al. [160] enhanced the SVM training process by integrating MFCC with Linear Discriminant Analysis (LDA) for more efficient dimensionality reduction, leading to an 86.44% accuracy rate in detecting voice pathologies within a subset of the SVD. Other common classifiers

[1]https://stimmdb.coli.uni-saarland.de/

used in the literature include artificial neural networks [169], hidden Markov models [9], and Gaussian mixture models [6].

Parallel to these efforts, advancements in end-to-end ML and deep learning techniques have further expanded the possibilities for VDD. Studies utilizing deep learning have mainly considered Convolutional Neural Network (CNN) models to extract acoustic features from spectrogram-based voice data automatically [144, 135] or in combination with hand-crafted features [49]. Nevertheless, the efficacy of these models can be compromised by limited dataset sizes, which often lead to overfitting and diminished performance on broader datasets.

An extensive exploration of audio adversarial attacks in the context of voice disorder detection systems has, to our knowledge, not yet been reported. Given the critical nature of healthcare applications, the potential implications of such adversarial interventions are manifold and particularly concerning. Firstly, introducing adversarial perturbations could lead to erroneous diagnoses, either by masking the presence of a disorder or falsely indicating its presence, thereby undermining the reliability of these systems in clinical settings. Furthermore, the integrity of patient data could be compromised, leading to privacy violations and loss of trust in digital healthcare solutions. The robustness of voice disorder detection models must be severely tested, exposing vulnerabilities that could be exploited to degrade system performance over time.

3.4 Contribution

Integrating ML into the diagnostic process offers an innovative solution that optimizes diagnostic times and precision. Computer-aided diagnosis systems have already shown promise in image processing for disease prediction [127] and can thus be applied to voice disorder detection, providing new methods to analyze acoustic data. However, the advancement of these technologies introduces potential risks that need careful examination. To address these concerns, this study examines the vulnerabilities of voice disorder detection systems to targeted manipulations used for side-channel attacks.

3.4.1 Vulnerabilities in Machine Learning-Based Voice Disorder Detection Systems

We investigate how such attacks can deliberately alter the system's output, causing it to erroneously classify *normal* voice samples as *pathological*. To assess different levels of vulnerability, we implement both simple attacks involving tone and pitch manipulation and more complex attacks, such as adversarial ones. For this pur-

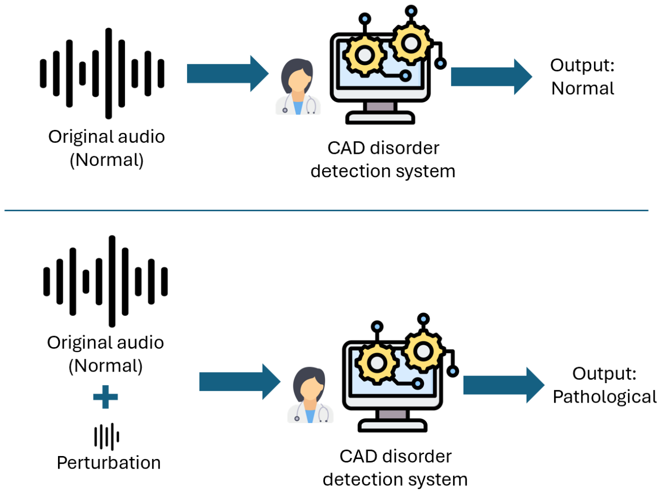


Figure 3.1: *High-level diagram of attacks for voice disorder detection systems.*

pose, we utilize a comprehensive approach encompassing *white-box* and *black-box* attack scenarios. The white-box attacks are conducted by leveraging two adversarial techniques based on gradients, exploiting complete visibility of the model's architecture and parameters. In parallel, black-box attacks are emulated by altering the pitch or adding an extraneous but imperceptible tone to the audio samples. This approach allows for a more realistic attack scenario (Fig. 3.1), where the adversary lacks direct knowledge of the model's inner workings and relies solely on its output to provoke misclassification. The proposed two-faceted approach ensures accurate and thorough investigations into the robustness of VDD systems against various adversarial tactics.

3.4.1.1 White-box attacks

In this study, we applied two well-known gradient-based techniques: Fast Gradient Sign Method (FGSM) [63] and Projected Gradient Descent (PGD) [110]. These methods explore the vulnerability of ML models to white-box adversarial attacks. FGSM computes the loss gradient with respect to the input features. It identifies the direction in which small perturbations can most effectively mislead the model. The gradient sign is then scaled by a value called epsilon (ϵ), which controls the intensity of the perturbation added to the original input. PGD extends FGSM by applying multiple small gradient updates iteratively to perturb the input data. The perturbation is adjusted within a predefined ϵ boundary, but iteratively, to find the optimal direction and magnitude [110].

Because attackers must know model architectures and parameters to compute gradients precisely, we define these as white-box attacks [18]

3.4.1.2 Black-box attacks

These attacks assume the attacker does not know the model's internal workings, including its architecture, weights, or the specific algorithms it employs. Instead, the attacker only has access to the model's input-output behavior [126, 15].

In this study, we applied two black-box evasion attacks: tone evasion and audio pitching. The first consists on inserting a specific tone into an audio sample. In this regard, it is possible to modify the tone parameters to be inserted (*e.g.*, in terms of amplitude, frequency, and phase) to balance the perturbation induced on the audio sample with the auditory perception of the disturbance. In our analysis, we selected several sine waves with varying frequencies and amplitudes. Specifically, we chose frequencies of 50, 75, 100, 125, and 150 Hz. For each frequency, we considered five different amplitudes to evaluate their auditory influence on the original audio: from causing a slight effect (0.2, 0.3, 0.4) to inducing a significant auditory disturbance (0.8, 0.9). Notably, higher amplitude values increase both the perturbation on the audio samples and the auditory perception of the disturbance [19].

A second attack proposed in this work is based on the substantial modification of the pitch of the audio samples. This experiment aims to determine if a voice disorder detection system is resilient enough to pitch modification on test samples. In this case, all the experiments were carried out by a pitch down of 5 steps (on a total of 12 per octave). This particular value is chosen to slightly modify the samples but leave all auditory properties as close to the original audio.

3.4.1.3 Datasets

For the experimental analysis, we used two different datasets, the HUPA dataset [195] and the Saarbruecken Voice Database (SVD)[2]. The HUPA dataset recorded by Universidad Politécnica de Madrid and Príncipe de Asturias Hospital of Alcalá de Henares is composed of two types of audio: 100 *normal* samples are related to users who do not have vocal difficulties, while 100 *pathol* samples are related to users with voice disorders (vocal fold polyps, nodules, edema, vocal leukoplakia, etc.). All audios are 1-3 seconds long and have 25 Hz sampling rate. The SVD database is an exhaustive collection of voice recordings by more than 2000 persons. It was collected at the Institute of Phonetics and Phoniatry, Caritas Clinic St. Theresia, Saarbrücken. The dataset comprises audio samples of the sustained vowel sounds /a/, /i/, and /u/ uttered at different intensities: normal, high, and low, along with variations in a rising and falling pitch pattern. Each sample was recorded at a frequency of 50 kHz. For the purposes of this study and for the sake of space, we focused exclusively on the vowels /a/ uttered at normal pitch. We then selected

[2] https://stimmdb.coli.uni-saarland.de/

three prevalent voice disorders: vocal fold cyst, vocal fold polyp, and unilateral vocal fold paralysis, to ensure a comparable sample size with the HUPA datasets. After this process, we obtained a subset of 520 recordings from individuals aged 16 and older, balanced with 260 normal and 260 pathological samples.

3.4.1.4 Voice disorder detectors: Implementation Details

For the implementation of voice disorder detectors, two different types of features were analyzed: (i) a low-level acoustic, mel-spectrogram feature, calculated as one spectrogram at mel-frequency from the Fourier spectrum using a nonlinear transformation at the frequency axis and normalized between 0 and 1; (ii) the Mel Frequency Cepstral Coeficients (MFCC) feature, obtained by applying Discrete Cosine Transform (DCT) to convert log Mel spectrum in the time domain.

Four different classifiers were subsequently used:

- The simple CNN proposed in [64], composed by two transposed 2D convolutional layers, a pooling layer, and multiple fully-connected layers.
- A CNN feature-extractor with an SVM classifier (linear kernel), as in [64].
- MobileNetV3Small [71, 189], a deep but at the same time fast network with low computational cost. We used a model pre-trained on natural images. To adapt the network and carry out fine-tuning, we replaced the last layer with a dense two-neuron layer and modified the first layer to manage a single-channel input. We used stochastic gradient descent as optimizer.
- MobileNetV3Small with an SVM classifier: in this case, the last dense layer of the MobileNet has been replaced with a Radial Basis Function SVM kernel.

Combining these feature extractors and classifiers, we obtained six different voice detectors that we used to evaluate the danger of the attacks created. To select the models on which to carry out the attack, each dataset was divided into training, validation and test (70%, 10%, 20%), and the best model was selected through 5-fold cross-validation.

Before extracting the features, each sample was subjected to a pre-processing phase to adapt to the input of the specific network. For MobileNet-based detectors, audio files were split into 200 ms snippets with an overlap of 160 ms (sampled at 25 kHz). For CNN-based detectors, each audio file is divided into multiple 1 s snippets with an overlap of 900 ms (sampled at 16 kHz).

Since each snippet represents a sample for the models, the evaluation produced both an accuracy on the snippets and on the entire audio file. The classification of the entire audio file was achieved by majority voting on the snippet predictions. In

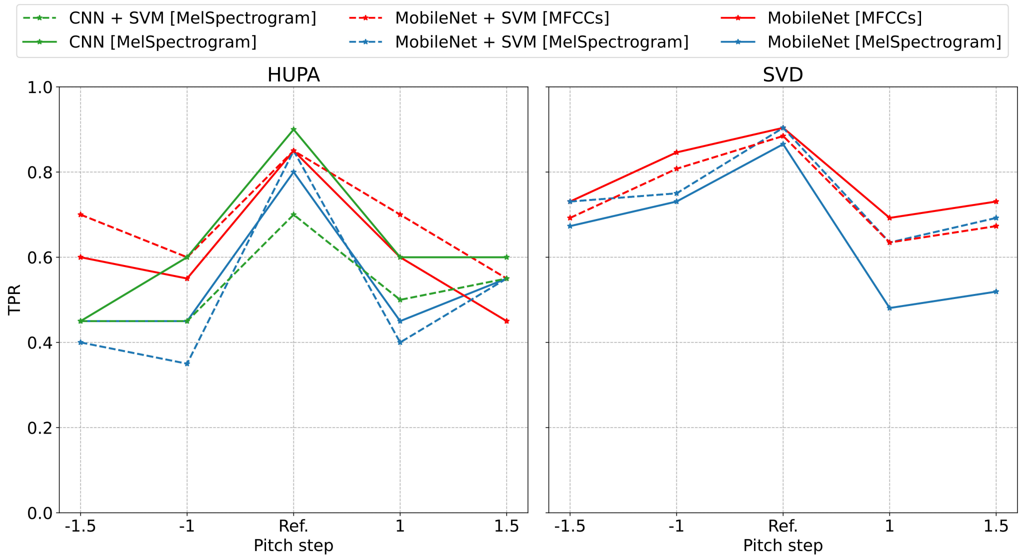


Figure 3.2: *Results of the pitch-based evasion attacks.*

particular, in the white-box scenario, since the attacker has knowledge of the entire architecture, he/she can attack the single correctly classified snippet. In the black box case, in order to obtain comparable results with the white box, we distinguish two scenarios: the attacker is completely uninformed, so he/she attacks the entire audio file, or the attacker is aware of the division into snippets and can thus attack a single snippet. Since the attacks were only carried out on correctly classified healthy samples, results are reported in terms of True Positive Rate (TPR).

3.4.1.5 Experimental Results

3.4.1.5.1 Black-box evaluation

Fig. 3.2 shows the results of the pitch-based evasion attacks. Changing the pitch in all cases leads to a decrease in TPR. However, it is also evident that the drop is less pronounced than in tone-based attacks. Except for the -1 step, the performance degradation appears generally uncorrelated with the pitch step values. The findings generally highlight the non-robustness of the models analyzed to normal variations in tone and pitch, which may occur not necessarily for attacks. In fact, although the mel-spectrogram features are more sensitive to these variations, all combinations of classifiers/features analyzed report a significant decrease in TPR.

The tone-based evasion attack results (Fig. 3.3, next page) show that adding a tone generally reduces the evaluated model's accuracy. Evaluations were carried out using tones at different frequencies and scales. It can be seen that the deterioration in accuracy is mainly related to the amplitude scale used for the attack.

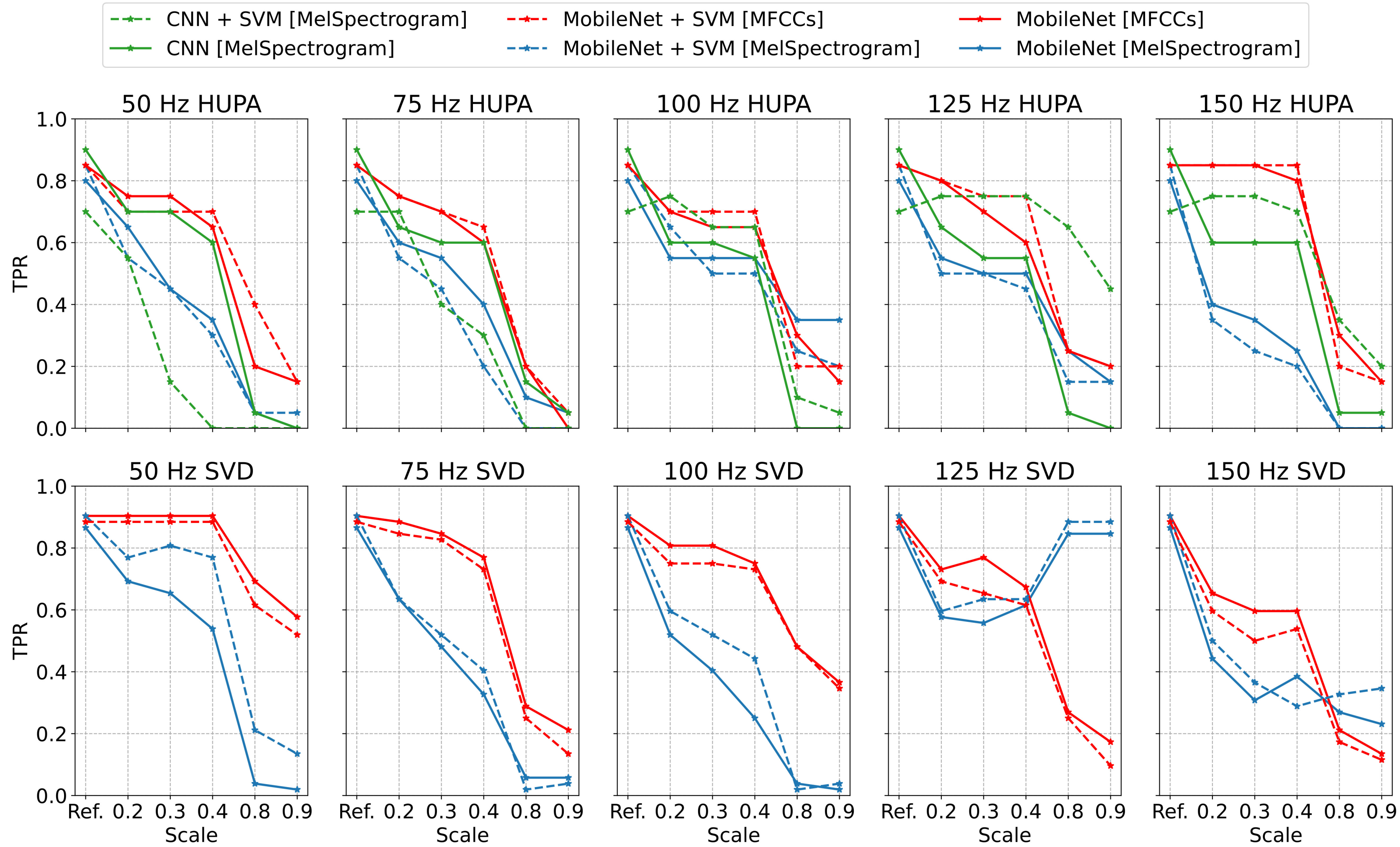


Figure 3.3: *Results of tone-based evasion attacks.*

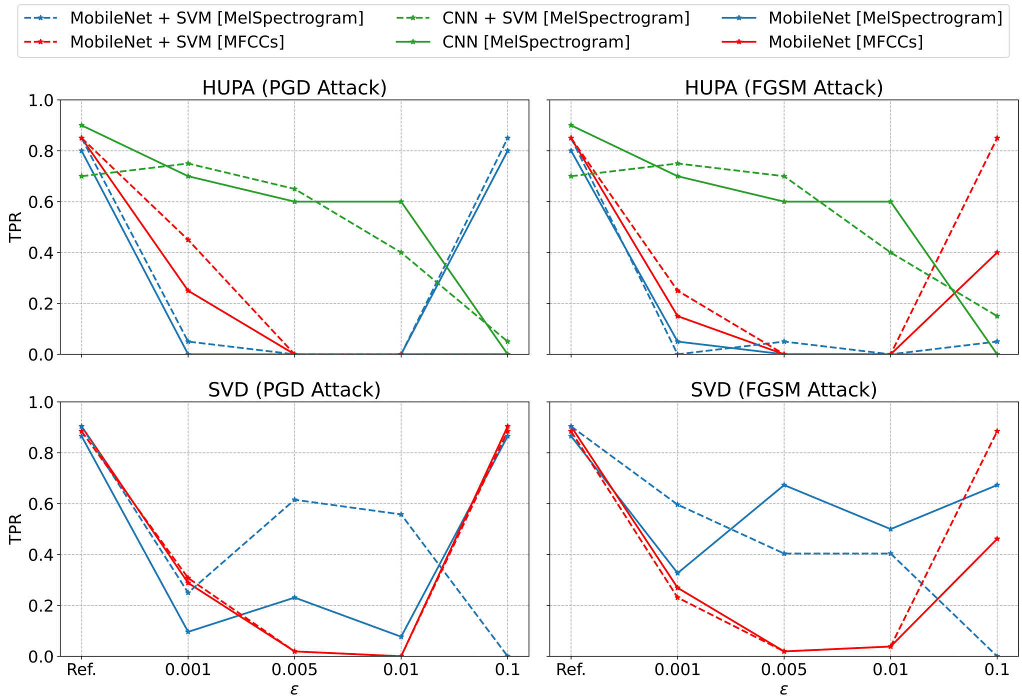


Figure 3.4: *Results of PGD and FGSM attacks, using different ε values.*

Higher tone scales result in more significant errors in classification: in particular, this manipulation causes the TPR, originally ranging between 70% and 90%, to drop to values below 40%. This is reasonable since higher-scale tones introduce more noise into the samples, leading to greater classification errors. On the other hand, by fixing the scale and varying the frequency from 50 to 150 Hz, less influence of the latter was seen in terms of classifier accuracy. Additionally, the results also show that overall the MFCC feature is more robust to this kind of manipulation.

3.4.1.6 White-box evaluation

In the white-box scenario (Fig. 3.4), both *PGD* and *FGSM* adversarial attacks demonstrated their effectiveness in deceiving the classifiers even with minimal perturbations ($\epsilon = 0.001$), especially when using a MobileNet architecture. This vulnerability is even more evident at higher epsilon values ($\epsilon = 0.1$), where the perturbation noise is severe enough to flatten the classifier's scores, either on the normal classification (FGSM) or on the pathological classification (PGD). On the other hand, CNNs show greater robustness against such attacks, maintaining better performance and stability across various perturbation levels. It is interesting to note the effectiveness of the *FGSM* attack, which, being much faster than *PGD*, can be used in real-time, deceiving the system without causing any slowdown.

As a final analysis, we computed the boxplots in Fig. 3.5 to illustrate the distribution of classifier scores for snippets of correctly classified files before and after attacks. For the sake of space, only the results of the classifiers based on the mel-spectrogram features on the HUPA dataset have been reported. However, they are representative of the overall trends observed. Each boxplot displays several key statistics: the median, the interquartile range (IQR, the box length which represents the spread of the middle 50% of the data), and potential outliers (points beyond the whiskers, which extend to 1.5 times the IQR from the quartiles). Overall, the original snippets exhibit lower median scores and narrower IQRs, indicating consistent and reliable classifier performance on clean data. In contrast, the attacked snippets show higher median scores and wider IQRs, highlighting the significant impact of perturbations on classification accuracy. Additionally, the presence of outliers in the attacked data suggests that the perturbations are causing extreme variations in the classifier's scores, further emphasizing the systems' vulnerability to adversarial attacks.

3.4.1.7 Discussion and conclusion

In this section, we analyzed the robustness of voice disorder detection systems to changes in input signals by evaluating both black-box and white-box attacks. Our results highlight a critical susceptibility of these systems to adversarial perturbations. To understand the extent of these perturbations on the original signals, we reported in Fig. 3.6 an example of original and manipulated signal for all attacks performed. While pitch and tone attacks introduce noticeable but specific alterations to the waveform, *FGSM* and *PGD* attacks, particularly at higher epsilon values, cause significant noise, demonstrating how these adversarial methods can distort the input signals and potentially compromise the detection systems.

The susceptibility of voice disorder detection systems to adversarial attacks poses a significant challenge to their reliability and effectiveness. Our future work will focus on developing robust defense mechanisms to enhance the resilience of these systems against such perturbations. A risk model will be associated with evaluating the concrete impact of neglecting true positives or caring for false positives with further medical treatments. In addition, the erroneous classification of normal audio as pathological can have significant ramifications outside clinical settings, as the case of cyberbullying or harassment campaigns.

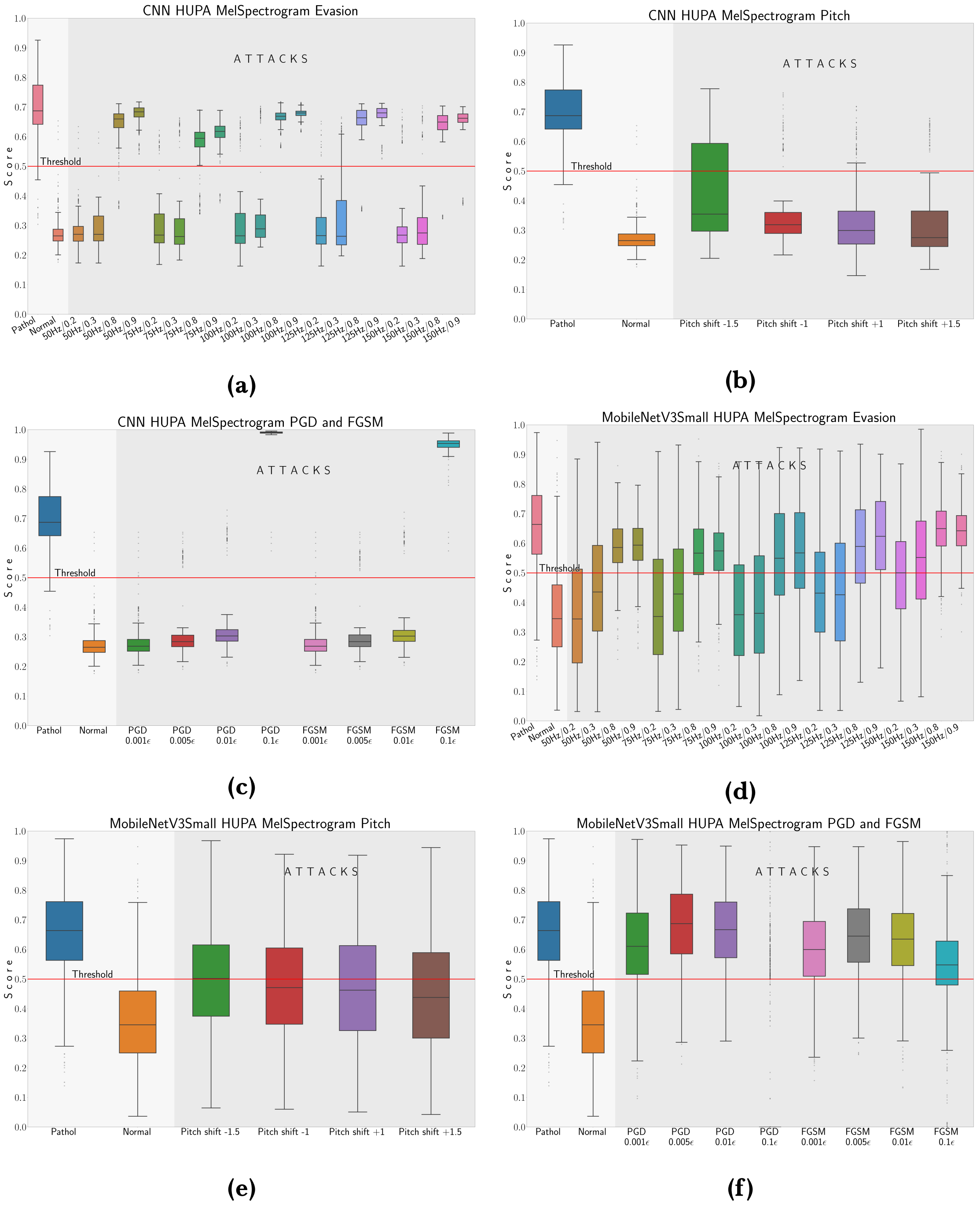


Figure 3.5: *Boxplots of scores for snippets of correctly classified files obtained with mel-spectrogram-based classifiers on the HUPA dataset. The boxplots compare the scores from the original unperturbed snippets (pathol and normal) with those subjected to black-box and white-box attacks.*

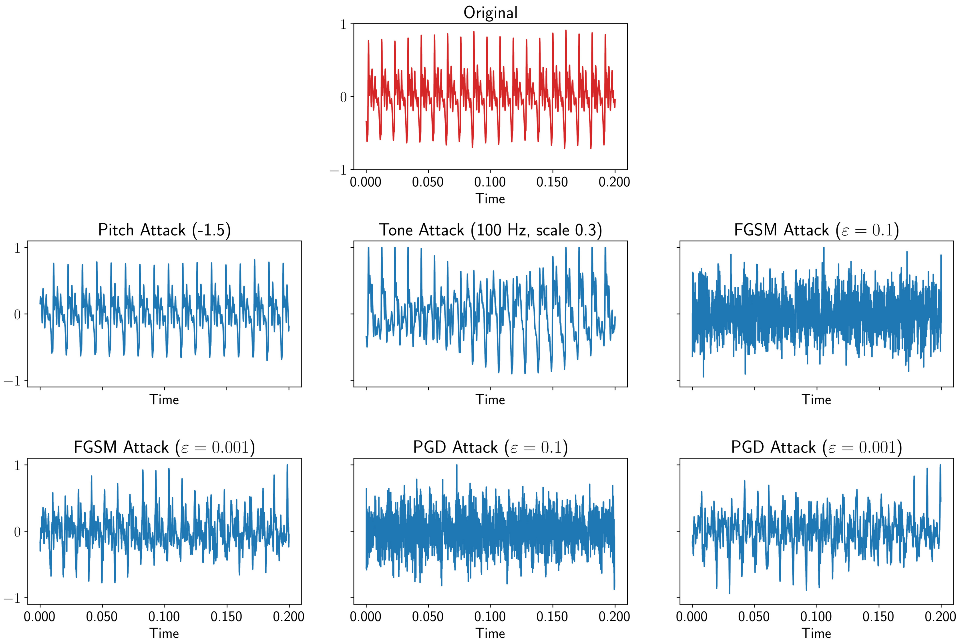


Figure 3.6: *Audio snippet subjected to different evasion attacks, both black box (tone and pitching) and white box (PGD and FGSM).*

Chapter 4

Privacy Risks In Human Activity Recognition Systems

In the previous chapter, an in-depth analysis of the vulnerabilities affecting ML models in voice disorder detection systems was presented. The focus now shifts from model weaknesses to the risks inherent in the data itself, which can be exploited through attacks that jeopardize user privacy. This chapter delves into privacy risk issues, examining how Human Activity Recognition (HAR) systems can inadvertently compromise users' privacy and the extent to which motion-based features can serve as biometric identifiers. After a brief introduction to the HAR systems, the discussion will move from the potential privacy risks to our contribution at the *IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering* (2025) with the paper titled *Unintended Identification in HAR Systems: Evaluating Privacy Risks.*

4.1 Introduction

Human Activity Recognition (HAR) systems, which utilize sensor data from mobile and wearable devices to classify physical movements, are increasingly integrated into daily life. These systems, leveraging embedded accelerometers, gyroscopes, and magnetometers, enable applications such as fitness tracking, healthcare monitoring, security, and behavioral analytics [141, 39, 122, 54, 145]. Despite their utility, HAR systems raise significant privacy concerns due to their processing of fine-grained motion data. Notably [37], HAR features can encode unique movement patterns, potentially enabling user or device re-identification [74]. This possibility allows attackers to track individuals over time, even with seemingly innocuous activity data [163].

4.2 Privacy and Data Protection in HAR Systems

Previous studies have primarily analyzed privacy risks in HAR systems under controlled scenarios. They typically presume that the adversary possesses prior knowledge of the user's activity [130]. Although these investigations provide a baseline for evaluating the discriminative capacity of HAR features, they do not represent realistic adversarial conditions. In real-world contexts, an adversary must first determine the performed activity before attempting identification. This requirement significantly increases the complexity of the attack.

Even under simplified conditions, findings from prior research indicate a wider issue: motion-based features derived from sensors can function as biometric signatures, revealing patterns unique to individuals or devices. As a result, HAR data collected for activity classification may unintentionally include personal information that enables user re-identification or profiling. This observation situates HAR within the broader discourse on biometric data protection. In this regard, recent surveys on privacy-enhancing technologies (PETs) for biometric recognition provide unified frameworks to categorize existing approaches, highlight their limitations, and relate them to fundamental data protection principles [115].

This intrinsic link between functionality and identifiability situates HAR systems within the broader domain of privacy and data protection regulation. The General Data Protection Regulation (GDPR), formally known as Regulation (EU) 2016/679 [55], provides the legal framework for protecting individuals with regard to the processing of personal data. It enforces strict requirements on how data are collected, stored, and processed, emphasizing transparency, fairness, and user control through principles such as consent and the *right to be forgotten*.

According to the GDPR, any processing of personal data, including biometric or behavioral data derived from HAR systems, must adhere to the principles of lawfulness, transparency, and purpose limitation. Data minimization is also a core requirement, ensuring that only information strictly necessary for the intended purpose is processed. HAR systems, however, often rely on weak biometric cues such as gait, motion dynamics, and even implicit location traces. These features are indispensable for reliable activity recognition, but they also carry inherent risks. Inadvertent disclosure could enable identity inference, behavioral profiling, or exposure to physical and digital threats.

Furthermore, centralizing sensitive data in large-scale infrastructure or cloud-based platforms magnifies potential vulnerabilities. A data breach or malicious exploitation could compromise not only individual privacy but also public trust in AI-driven health and behavioral monitoring technologies. Therefore, addressing privacy risks in HAR systems is not merely a technical challenge; it is a fundamental requirement for ensuring the ethical and lawful deployment of AI in everyday life.

4.3 From HAR to User Re-identification

HAR has seen rapid development in recent years, primarily due to mobile and wearable sensors. The proliferation of sensors has resulted in several applications, many of which have been tested with inertial sensors (accelerometers and gyroscopes), resulting in advantages and reliability across diverse uses [141, 122]. HAR applications' ability to classify activities such as walking, running, and sitting in real time makes them essential assets across various fields. Moreover, the integration of HAR into smartphone platforms and IoT ecosystems has broadened their use, making activity-aware services increasingly accessible and scalable [162].

However, alongside these benefits, emerging studies have raised concerns about the unintended consequences of using motion data. In particular, researchers have highlighted that the same HAR features used to classify activities can also be repurposed for identifying individuals. As motion patterns tend to be influenced by physiological and behavioral traits, HAR data often carries unique signatures tied to the user, effectively performing as a form of soft biometrics [74, 34]. This opens up potential avenues for user re-identification and tracking, even in the absence of explicit identity labels.

Similar concerns have been raised in the domain of face recognition, where embeddings not only enable identity verification but also expose soft-biometric attributes such as gender or age. To mitigate this leakage, recent work introduced Multi-IVE, a privacy-enhancing method that extends Incremental Variable Elimination (IVE) to suppress multiple demographic attributes simultaneously from face embeddings [117]. This parallel suggests that HAR embeddings, much like face embeddings, may also encode sensitive personal traits beyond activity classification, indicating the need for exploring privacy-preserving approaches in motion-based biometrics.

Several works have demonstrated the likelihood of this threat. For instance, identification frameworks based solely on accelerometer and gyroscope data have achieved accuracies close to 98–100% in controlled environments [8]. Even when individuals perform the same activities, their movement dynamics can be differentiated with high precision. More concerning, adversaries can often infer user identities from sensor data even when anonymized or recorded under different activity contexts, as long as sufficient motion traces are available [150].

In [162], the authors confirmed that de-identified activity data from smartphones and wearables can often be re-linked to individuals, with re-identification rates exceeding 86% in many cases. These findings underscore a critical risk: although HAR systems are not inherently designed for identity recognition, they may still encode sufficient biometric information to compromise user privacy.

In a real-world scenario, however, the possible attacker does not know in advance what activity the target user was performing in the captured sensor data. In this regard, the authors of [150] propose a deep re-identification model evaluated on wrist-worn accelerometer data collected in the wild. In their threat model, an attacker obtains an anonymized sensor dataset without knowledge of the underlying activities. To simulate this, they employ a prior HAR stage: an activity classifier is used to automatically label segments of the sensor streams (*e.g.*, walking, running, stationary) before attempting identification. They demonstrate high re-identification rates even when the attacker has no prior knowledge of the victim's actions, highlighting a critical privacy vulnerability in motion data sharing. Another line of research avoids assuming any fixed activity during identification by leveraging multi-task learning and activity-agnostic feature representations. In [156], the authors developed a weakly-supervised Siamese network that jointly learns to recognize both who the user is and what activity they are doing, without requiring strong labels for either task.

While most studies focus on demonstrating the feasibility of re-identification from motion data, a smaller line of research has started to address these risks through privacy-enhancing methods. For instance, the *GaitPrivacyON* framework introduces a privacy-preserving mobile gait verification approach, where convolutional autoencoders transform raw gait signals into representations that conceal sensitive attributes (*e.g.*, gender, activity) before user verification. This approach achieves strong authentication performance while significantly reducing privacy leakage, highlighting a possible path to reconcile HAR-based biometrics with privacy protection [38].

Beyond multitask learning, researchers have also explored *activity-free* identification methods that aim for continuous user recognition during natural behavior. In the article [151], the authors introduce an identification technique that does not require any specific gesture or pre-defined activity from the user. The proposed approach transforms raw accelerometer and gyroscope streams into image-like representations (*e.g.*, time–frequency spectrograms) and applies deep convolutional neural networks inspired by computer vision. By not relying on any particular activity, this method can passively identify users *in the wild* as they go about their daily routines.

Other recognition pipelines rely on multiple activities of daily living. For instance, in [114], a deep residual neural network for smartwatch-based user identification leverages diverse activity data. Rather than training on a single motion (*e.g.*, only gait), they incorporate a range of daily activities (walking, sitting, using

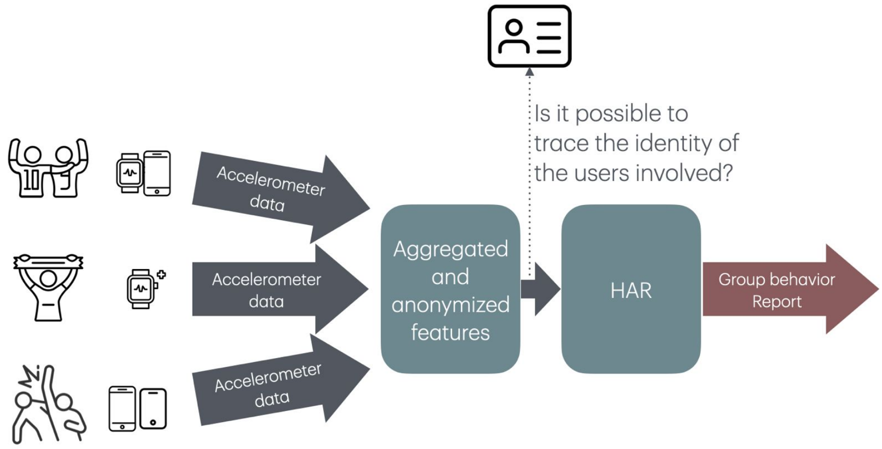


Figure 4.1: *Overview of the study's objective: evaluating if it is possible to trace the identity of individuals involved in HAR.*

a phone, etc.) into the model's learning process. This yields a more generalized user profile from inertial patterns.

While prior work has demonstrated the feasibility of user re-identification from motion data, even under activity-agnostic or anonymized conditions, existing studies often rely on complex deep learning models, multi-sensor setups, or focus on specific device types. In contrast, the next section provides a systematic and interpretable assessment of privacy risks in HAR systems using only accelerometer data, structured over three realistic identification scenarios.

4.4 Contribution

Much of the existing literature relies on multiple sensors (e.g., accelerometer, gyroscope, magnetometer), leaving open the question of whether a single sensor, specifically the accelerometer, is sufficient for user or device differentiation. This section investigates whether accelerometer data alone enables re-identification, thus providing insight into privacy risks associated with minimal sensor information. The accelerometer, due to its ubiquity in HAR systems, serves as a critical test case. If user or device identification proves feasible using only accelerometer data, this finding underscores the significant privacy implications of even limited motion-based sensor data. In this section, we introduce realistic experimental protocols by incorporating HAR-based activity inference before performing user or device identification. This better reflects real-world attack scenarios, where an adversary lacks prior knowledge of the performed activity (Fig. 4.1).

We structure our analysis around three key cases:

- Intra-activity identification: Evaluating if different samples of the same activity are sufficient to distinguish users.
- Cross-activity identification: Exploring whether identity traces persist across different activities.
- HAR-guided identification: Simulating a realistic scenario where activity recognition is performed first, and identification is then attempted on similar data.

4.4.1 Unintended Identification in HAR Systems: Evaluating Privacy Risks

The methodology adopted in this study aims to evaluate whether accelerometer signals enable unintended user identification systematically. As briefly anticipated, we design three experimental scenarios corresponding to different privacy threats:

- *Intra-activity* identification: We evaluated whether samples of the same activity could distinguish individuals. For each sample, we compared it to other samples from the same activity, determining whether the closest match was genuine (same user) or an impostor (different user).
- *Cross-activity* identification: This scenario assessed whether identity-specific patterns persist across different activities. Here, we compared samples from one activity against samples from different activities, again distinguishing genuine from impostor matches.
- *HAR-guided* identification: To simulate a realistic identification scenario, we first predicted the activity of each sample using a HAR classifier trained on ideal conditions and then applied it to real-world conditions. After predicting the activity, samples were compared within their predicted activity class to identify genuine or impostor matches.

4.4.1.1 Datasets

We employ two public HAR datasets that provide real-life information: Gonzalez [53] and WISDM [92, 108] dataset.

The Gonzalez dataset is a publicly available collection explicitly designed for real-life HAR using smartphone sensors. It was collected through an Android application that recorded data from accelerometer, gyroscope, magnetometer, and GPS sensors. A total of 19 participants performed various activities categorized into four distinct classes: inactive, active, walking, and driving. The dataset ensures independence from device orientation, placement, and individual characteristics. Each participant recorded multiple sessions corresponding to different activities, resulting in approximately 680,000 seconds of data and 29 million sensor measurements. In this study, we focused exclusively on accelerometer data. The 19 participants were used to test the personal recognition system.

The WISDM dataset consists of accelerometer data collected from Android smartphones carried by users performing common daily activities such as walking, jogging, ascending/descending stairs, sitting, and standing. The dataset comprises two versions: *Activity Prediction*, an ideal version acquired under controlled conditions (14 users), and *Actitracker*, a in-the-wild version gathered in real-world, uncontrolled conditions (225 users). The users carried the devices in their pants pockets, with data recorded at a sampling rate of 20 Hz (every 50 ms). We use the WISDM Activity Prediction dataset for training and testing the HAR system. The WISDM Actitracker is used to test the personal recognition system.

Regarding the features used, we followed the approach in [53], so we did not use raw sensor data but rather features that were particularly discriminating for HAR tasks, considering the context in which HAR could operate. Therefore, we extracted very simple time-domain features from the accelerometer data (x, y, z measurements), using a window size of 20 seconds with a 19-second overlap:

- *Mean*: The average value of the sensor readings over a specific time window.
- *Variance*: A measure of the variability of the sensor readings, indicating the extent of movement.
- *Median Absolute Deviation (MAD)*: A robust measure of variability that is less sensitive to outliers.
- *Maximum and Minimum*: The extreme values of the sensor readings within a time window, indicating the range of motion.
- *Interquartile Range (IQR)*: The range between the first and third quartiles, providing a measure of statistical dispersion.

This configuration was chosen based on its effectiveness in previous HAR studies and its ability to capture both short-term and long-term activities. Following feature extraction, Z-score normalization was applied to each feature dimension across all data samples.

4.4.1.2 Experimental Protocol

The experimental protocol was meticulously developed to comprehensively assess the discriminative power of commonly used Human Activity Recognition (HAR) features within a framework analogous to contemporary biometric systems. The overarching objective of this study is to provide a deeper, more systematic investigation of whether features extracted from accelerometer data can reliably and robustly distinguish among individuals across a variety of realistic scenarios. By extending the scope of previous studies, this work not only evaluates the separability of individuals based on sensor-derived features but also critically examines the broader implications for user privacy and the potential risks posed by the use of such features in pervasive computing environments. This research aims to bridge the gap between HAR-based identification and established biometric evaluation protocols, thereby laying a foundation for future developments in privacy-aware mobile sensing.

To ensure a fair, unbiased, and robust evaluation applicable across all experimental scenarios, a consistent identification protocol was meticulously established. User-specific templates were constructed using data derived from dedicated participant sessions, while test data were strictly drawn from separate, non-overlapping sessions that were not utilized in the creation of templates. This approach emulates best practices in biometric system evaluation, maximizing generalizability and minimizing overfitting or session-specific bias. The identification process itself entailed a comprehensive, pairwise comparison: each target user's template was systematically compared against all available test records, encompassing both mated comparison trials (where the same individual is present in both template and probe) and non-mated comparison trials (involving different individuals), as rigorously defined by the ISO standard [76]. To quantify similarity, the classical Euclidean distance metric was used to compare feature vectors, providing an interpretable and widely accepted baseline for subsequent analysis.

Building upon this general identification framework, the protocol was systematically adapted to address the unique challenges and objectives posed by each of the defined experimental scenarios. These scenario-specific adaptations are described in detail below:

- In the *intra-activity* identification scenario, the aim was to distinguish individuals performing the same type of activity. Consequently, for a given sample, identification comparisons (genuine and impostor trials) were made exclusively against samples corresponding to the same activity.

- For the *cross-activity* identification scenario, the focus was on assessing whether identity-specific patterns persisted between different types of activity. In this case, identification comparisons for each sample of one activity were made against other samples explicitly chosen from different activities.

- For the *HAR-guided* identification scenario, a two-stage process was followed:

1. Activity prediction: first, the HAR classifier was developed by training a Support Vector Machine (SVM) featuring a Radial Basis Function (RBF) kernel. This model was trained using the *controlled* version of the WISDM dataset (Activity Prediction) [92]. Subsequently, this trained HAR classifier was used to predict the activity class for each probe sample of the identity verification system.

2. Conditional identity verification: after the HAR system assigned an activity label to each probe sample, identity verification was performed only within that activity class. In other words, each probe was compared with samples from the same predicted activity, restricting both genuine and impostor comparisons to that class. When applying our method to the Gonzales et al. dataset, we mapped the HAR system's output to our target classes due to the limited set of activity labels available (walking, active, driving, inactive). Specifically, walking, jogging, and stairs were grouped under the *WALKING* category, while sitting and standing were grouped as *ACTIVE*. In contrast, for the WISDM Actitracker dataset, no mapping was required since the HAR output classes directly matched the activity labels of the probe samples.

   It is important to acknowledge that the HAR accuracy observed on the Gonzales et al. dataset (75%) and the WISDM Actitracker dataset (63%) is relatively modest, owing to the inherent challenges of cross-dataset evaluation. Nevertheless, this limitation does not undermine the study's primary aims. The focus is not on maximizing activity recognition performance, but rather on simulating realistic application scenarios in which imperfect recognition may occur. The central research question is to what extent the integration of activity information, even when noisy or suboptimal, can enhance the process of individual re-identification, thereby illuminating the interplay between HAR accuracy and biometric distinctiveness in practical deployments.

Across all experimental protocols, a critical methodological principle was maintained: templates and probes were always selected from different sessions to rigorously simulate realistic recognition conditions and ensure the validity of the evaluation. In the case of the Gonzales dataset, these sessions correspond to original, naturally segmented recordings with an average duration of approximately 20 minutes, each acquired during distinct recording phases to capture day-to-day variability. Conversely, for the WISDM dataset, where only a single continuous recording per participant was available, a novel sessionization strategy was devised. The continuous data stream was partitioned into multiple artificial sessions by segmenting the recording into chunks of up to 5 minutes, separated by a 2-second interval. This approach, while artificial, was carefully designed to introduce temporal separation and to emulate better the multi-session conditions typically found in real-world deployments, thereby enhancing the ecological validity of the experimental results. The number of users considered varies depending on the protocol and the specific activity analyzed, as not all users performed every activity or have recordings of equal duration. In particular, users without enough data to support multiple sessions are excluded from the analysis. The exact number of users included in each case is reported in Table 4.1 for the Garcia-Gonzales et al. dataset and in Table 4.2 for the WISDM dataset.

Table 4.1: *Number of users for intra-activity, cross-activity, and HAR-guided identification using the Garcia-Gonzales et al. dataset.*

Activity/EER	Intra-activity	Cross-activity	HAR-guided
Active	9	10	9
Driving	10	12	-
Inactive	10	12	-
Walking	13	14	13

Table 4.2: *Number of users for intra-activity, cross-activity, and HAR-guided identification using the WISDM dataset.*

Activity	Intra-activity	Cross-activity	HAR-guided
Standing	15	11	15
Stairs	4	5	3
Sitting	28	25	28
Jogging	34	37	34
Walking	62	48	64

The results of this analysis are reported in terms of the False Match Rate (FMR) and Genuine Acceptance Rate (GAR). FMR represents the proportion of non-mated comparison trials where the system incorrectly decided on a match between a test record and a template from different individuals. GAR, on the other hand, indicates the percentage of mated trials where the system correctly matched the test record to the template of the same individual. While a high GAR is typically desirable for accurate recognition, in this context, it also raises privacy concerns, as it implies the system's potential to identify individuals based on the HAR features.

To comprehensively evaluate the system's performance, Receiver Operating Characteristic (ROC) curves are also reported, illustrating the trade-off between the FMR and GAR across different decision thresholds.

Table 4.3: *EER for intra-activity, cross-activity, and HAR-guided identification using the Garcia-Gonzales et al. dataset.*

Activity/EER	Intra-activity	Cross-activity	HAR-guided
Active	44.01%	48.44%	48.04%
Driving	41.48%	46.67%	-
Inactive	22.36%	42.88%	-
Walking	39.25%	46.81%	41.80%

Table 4.4: *EER for intra-activity, cross-activity, and HAR-guided identification using the WISDM dataset.*

Activity	Intra-activity	Cross-activity	HAR-guided
Standing	27.60%	24.86%	38.28%
Stairs	14.12%	28.88%	10.80%
Sitting	18.29%	39.37%	18.09%
Jogging	27.81%	36.88%	27.47%
Walking	27.60%	41.58%	32.60%

4.4.1.3 Experimental results

The experimental results summarized in Tables 4.3 and 4.4 highlight several aspects related to accelerometer-based user identification across different scenarios. In the *intra-activity* scenario, identification performance varies significantly across activity types and datasets. Tasks involving limited motion, such as *Inactive* in the Garcia-Gonzales dataset, or *Standing* and *Sitting* in WISDM, tend to produce lower error rates, likely due to the influence of stable sensor-related characteristics rather than distinctive user-specific traits. The *Stairs* activity in WISDM also shows a particularly low EER (14.12%), but the result is difficult to interpret reliably given that it is based on only four users. In contrast, more dynamic activities like *Active* or *Jogging* lead to higher identification errors, as greater movement variability reduces the consistency of motion patterns.

The performance difference between the two datasets can be attributed to the nature of data acquisition and preprocessing. In the case of the Gonzales dataset, the signals are pre-filtered at the source, potentially leading to the loss of information useful for user recognition. Moreover, data are collected across different acquisition phases. In contrast, the WISDM dataset consists of a single continuous signal acquired sequentially and later segmented by us into multiple sections, preserving more raw information and temporal consistency. For these reasons, the Gonzales dataset may be more representative of a real-world scenario. However, a HAR system might expose unfiltered data or data collected under acquisition conditions very similar to those of the enrollment phase. In this sense, the WISDM dataset can be considered a *worst-case* scenario for data exposure and the associated risk of user re-identification.

In the cross-activity setting, the substantial increase in error rates indicates a lack of generalization of user-specific motion cues across tasks. The relationship between movement patterns and identity appears to be strongly dependent on the context in which the data is collected.

In the HAR-guided scenario, where identification follows a preliminary activity classification step, the results remain largely in line with those of the intra-activity setting. A slight degradation is observed in most cases, reflecting the impact of error propagation from the HAR stage. Despite the limited changes in overall performance, this setting deserves particular attention, as it realistically reflects how an adversary could operate in practice, first inferring the activity, then attempting re-identification, demonstrating that the attack remains feasible.

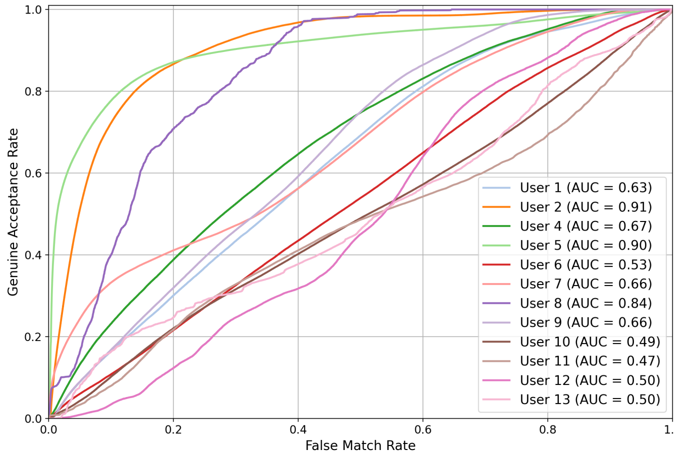
Individual-level ROC analyses (Figs. 4.2 and 4.3) reveal considerable variability in user identifiability across both datasets. In the Garcia-Gonzales set, some individuals exhibit highly distinctive motion patterns, with AUC values reaching 0.91 in the *intra-activity* setting and remaining elevated even in the *cross-activities* (e.g., User 2, $AUC = 0.79$). In WISDM, several users achieve near-perfect scores in the intra-activity case ($AUC \geq 0.99$) and retain high identifiability under *cross-activity* and *HAR-guided* protocols. Such results confirm that, despite high global error rates, certain users remain persistently exposed to re-identification, highlighting privacy risks that are not apparent from average performance alone.

In summary, user identification based on accelerometer-derived HAR features shows moderate overall effectiveness, with performance strongly influenced by activity type and inter-user variability. The presence of consistent individual-level vulnerabilities calls for privacy protection strategies that are both scenario-aware and sensitive to user-specific risk profiles.

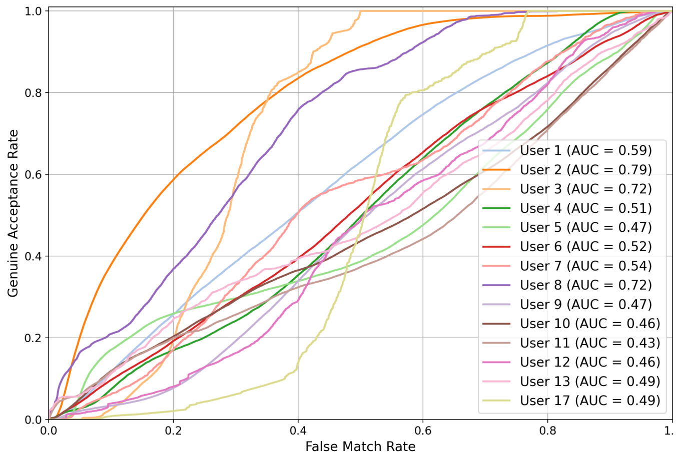
4.4.1.4 Conclusions

The analysis reported in this work aimed to evaluate the discrimination power of some HAR features obtained from mobile device accelerometers, in particular, whether they can be used for personal recognition. This investigation confirms that even basic accelerometer HAR features (mean, variance, MAD, minimum, maximum, and IQR) can lead to unintended user or device re-identification across a range of realistic threat models. Our experiments show that, although overall error rates were moderate, specific users remained consistently and highly identifiable, and the HAR-guided pipeline demonstrated that an adversary who first infers activity labels can still achieve re-identification performance comparable to an idealized *intra-activity* setting. These findings highlight a crucial privacy issue:

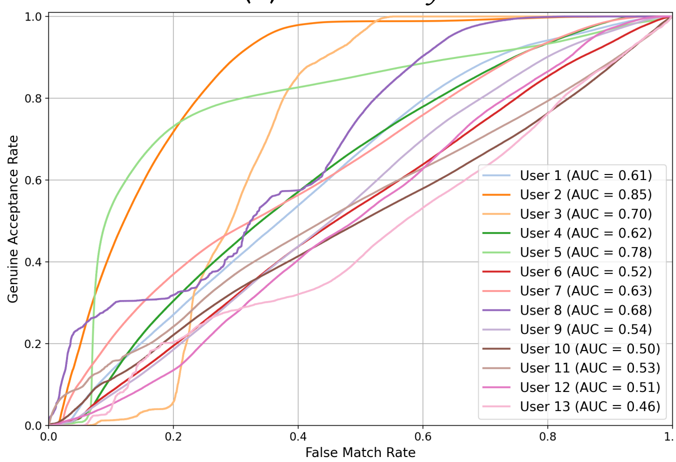
motion data that are routinely shared for harmless analytics or health purposes may expose individuals to the risk of user or device identification, even when only a single sensor is involved. In the current data-driven landscape, addressing this vulnerability is crucial to preserving privacy and ensuring that the benefits of HAR sensing do not come at the cost of trust.



(a) *Intra-activity*

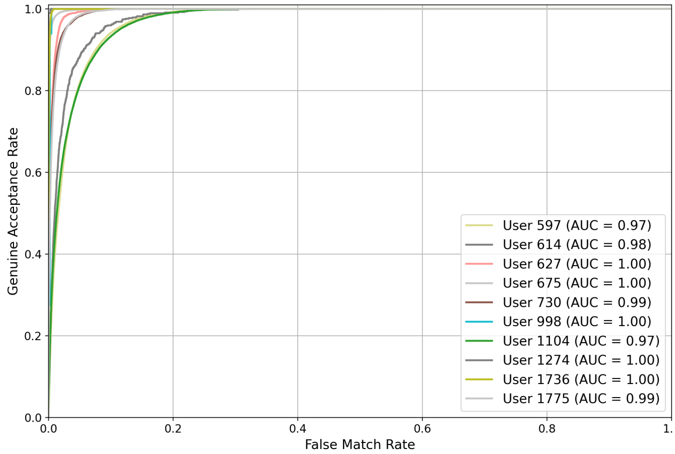


(b) *Cross-activity*

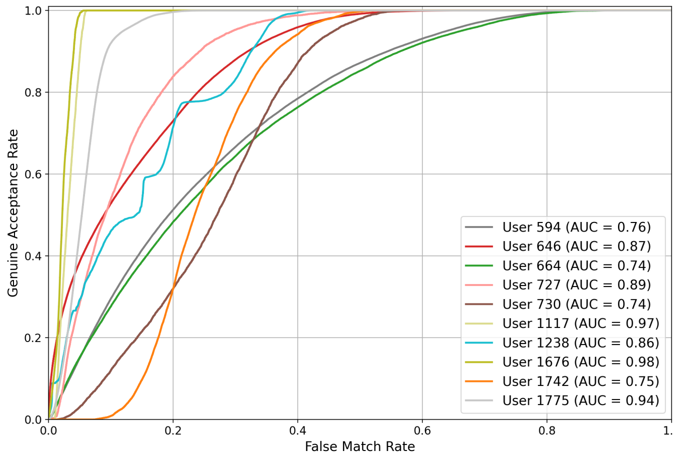


(c) *HAR-guided*

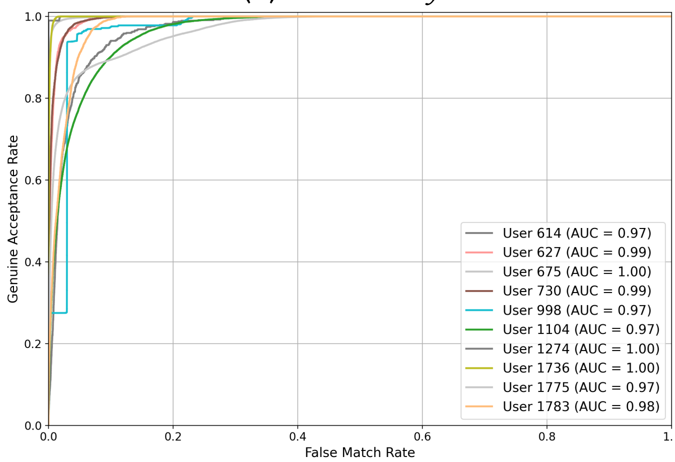
Figure 4.2: *ROC curves related to Walking activity for individual users using the Garcia-Gonzales et al. dataset across different protocols: (a) Intra-activity, (b) Cross-activity, (c) HAR-guided.*



(a) *Intra-activity*



(b) *Cross-activity*



(c) *HAR-guided*

Figure 4.3: *ROC curves related to Walking activity for individual users using the WISDM dataset across different protocols.*

Chapter 5

General Conclusions and Research Perspectives

This thesis explored the phenomenon of deepfakes from both technological and defensive perspectives. It analyzed the foundations and evolution of deepfake generation methods and provided a structured classification of the main manipulation categories.

The research then focused on detection, emphasizing the challenges posed by compression and proposing two novel methods to improve robustness under these conditions. The first approach leverages frequency representations through the Discrete Cosine Transform (DCT), while the second incorporates compression-aware information directly into the network architecture, introducing the Multi-Head and Multi-Branch architectures. In addition, two augmentation strategies specifically designed for deepfake video detectors were presented: a frame-centric approach and a video-centric one.

The results demonstrate that encoding frequency and compression information within a DCT-based tensor representation can substantially enhance detection reliability in real-world scenarios, where media are often degraded by social network transmission. Furthermore, appropriately designed augmentation strategies for video data can further improve detection accuracy. The proposed Multi-Head and Multi-Branch architectures also proved effective as general boosters for several existing deepfake detectors.

Although the experimental analysis was limited to a few datasets due to computational and practical constraints, the study provides a strong foundation for enhancing existing detection models and guiding future research in this field.

Beyond deepfakes, the thesis also investigated vulnerabilities in machine learning systems applied to voice disorder detection. It introduced a series of white-box and black-box adversarial attacks that, to the best of our knowledge, had not yet been reported in the state of the art. The results highlight the urgent need for robust countermeasures to ensure the reliability and security of such systems.

Finally, potential privacy risks associated with data used in HAR systems were examined. The analyses revealed the possibility of user re-identification even when limiting the input to a single sensor (the accelerometer). Three identification scenarios were explored (intra-activity, cross-activity, and HAR-guided), demonstrating that privacy risks persist across diverse contexts.

Overall, this work contributes to maintaining a high level of awareness regarding artificial intelligence systems, which, despite achieving remarkable results, can sometimes be seriously deceived. The thesis provides a comprehensive overview of the technological advances, vulnerabilities, and risks associated with modern AI-based systems, while outlining possible directions for future research toward more secure, explainable, and trustworthy models.

While this work shows promising results, there are still several areas for further research. Future studies should focus on improving the generalization capabilities of deepfake detectors not only across different compression levels but also across different algorithms, especially those common on social media platforms. From a security perspective, further research is needed to develop effective defenses against adversarial attacks targeting both multimedia and medical AI systems. Finally, the privacy risks identified in HAR systems require the exploration of privacy-preserving learning paradigms that can balance performance and user anonymity. Addressing these challenges represents a crucial step towards implementing more robust, secure, and reliable AI systems.

Bibliography

[1] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 01 2019.

[2] Google AI. Contributing data to deepfake detection research, 2019.

[3] Fahad Taha Al-Dhief, Marina Mat Baki, Nurul Mu'azzah Abdul Latiff, Nik Noordini Nik Abd Malik, Naseer Sabri Salim, Musatafa Abbas Abbood Albader, Nor Muzlifah Mahyuddin, and Mazin Abed Mohammed. Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access*, 9:77293–77306, 2021.

[4] Ahmed Al-Nasheri, Ghulam Muhammad, Mansour Alsulaiman, Zulfiqar Ali, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, and Mohamed A Bencherif. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, 31(1):113–e9, 2017.

[5] Abdullah Alanazi. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924, 2022.

[6] Zulfiqar Ali, Ghulam Muhammad, and Mohammed F Alhamid. An automatic health monitoring system for patients suffering from voice complications in smart cities. *IEEE Access*, 5:3900–3908, 2017.

[7] Fahd Saleh Alotaibi. Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, 10(6), 2019.

[8] Abdulrahman Alruban, Hind Alobaidi, and Nathan Clarke'Fudong Li. Physical activity recognition by utilising smartphone sensor signals. *arXiv preprint arXiv:2201.08688*, 2022.

[9] Redouane Benhammoud and Abdellah Kacha. Automatic classification of disordered voices with hidden markov models. In *Proceedings of the 2018 International Conference on Signal, Image, Vision and their Applications (SIVA)*, pages 1–6. IEEE, 2018.

[10] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

[11] Taís Fernanda Blauth, Oskar Josef Gstrein, and Andrej Zwitter. Artificial intelligence crime: An overview of malicious use and abuse of ai. *Ieee Access*, 10:77110–77122, 2022.

[12] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training strategies and data augmentations in cnn-based deepfake video detection. In *Proceedings of the 2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020.

[13] DR Boone. The voice and voice therapy. *Allyn and Bacon google schola*, 2:830–843, 2005.

[14] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[15] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ArXiv*, abs/1712.04248, 2017.

[16] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4103–4112, June 2022.

[17] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2017.

[18] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks . In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, Los Alamitos, CA, USA, May 2017. IEEE Computer Society.

[19] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *Proceedings of the 2018 IEEE security and privacy*

workshops (SPW), pages 1–7, Los Alamitos, CA, USA, May 2018. IEEE, IEEE Computer Society.

[20] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection, 2021.

[21] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7196–7205, 2021.

[22] Guan-Lin Chen and Chih-Chung Hsu. Jointly defending deepfake manipulation and adversarial attack using decoy mechanism. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9922–9931, Aug 2023.

[23] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1081–1088, 2021.

[24] Saheb Chhabra, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Low quality deepfake detection via unseen artifacts. *IEEE Transactions on Artificial Intelligence*, pages 1–13, 2023.

[25] Yunjey Choi, Min-Je Choi, Mun Su Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2017.

[26] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017.

[27] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Honolulu, HI, July 2017. IEEE.

[28] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Sscgan: Facial attribute editing via style skip connections. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the Computer Vision – ECCV 2020*, pages 414–429, Cham, 2020. Springer International Publishing.

[29] Jacob Cohen. *Statistical power analysis for the behavioral sciences.* routledge, 2013.

[30] Seth M Cohen, Jaewhan Kim, Nelson Roy, Carl Asche, and Mark Courey. Direct health care costs of laryngeal diseases and disorders. *The Laryngoscope*, 122(7):1582–1588, 2012.

[31] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi. Tensor-based deepfake detection in scaled and compressed images. In *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)*, pages 3121–3125, 2022.

[32] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2019.

[33] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Erhan Davarci and Emin Anarim. User identification on smartphones with motion sensors and touching behaviors. In *Proceedings of the 2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2022.

[35] Audrey de Rancourt-Raymond and Nadia Smaili. The unethical use of deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

[36] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Luis F. Gomez, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Zhizhou Zhong, Yuge Huang, Yuxi Mi, Shouhong Ding, Shuigeng Zhou, Shuai He, Lingzhi Fu, Heng Cong, Rongyu Zhang, Zhihong Xiao, Evgeny Smirnov, Anton Pimenov, Aleksei Grigorev, Denis Timoshenko, Kaleb Mesfin Asfaw, Cheng Yaw Low, Hao Liu, Chuyi Wang, Qing Zuo, Zhixiang He, Hatef Otroshi Shahreza, Anjith George, Alexander Unnervik, Parsa Rahimi, Sébastien Marcel, Pedro C. Neto, Marco Huber, Jan Niklas Kolf, Naser Damer, Fadi Boutros, Jaime S. Cardoso, Ana F. Sequeira, Andrea Atzori, Gianni Fenu, Mirko Marras, Vitomir Štruc, Jiang Yu, Zhangjie Li, Jichun Li, Weisong Zhao, Zhen Lei, Xiangyu Zhu, Xiao-Yu Zhang, Bernardo Biesseck, Pedro Vidal, Luiz Coelho, Roger Granada, and David Menotti. Second frcsyn-ongoing: Winning solutions and post-challenge analysis to improve face recognition with synthetic data. *Information Fusion*, 120:103099, 2025.

[37] Paula Delgado-Santos, Giuseppe Stragapede, Ruben Tolosana, Richard Guest, Farzin Deravi, and Ruben Vera-Rodriguez. A survey of privacy vulnerabilities of mobile device sensors. *ACM Comput. Surv.*, 54(11s), September 2022.

[38] Paula Delgado-Santos, Ruben Tolosana, Richard Guest, Ruben Vera-Rodriguez, Farzin Deravi, and Aythami Morales. Gaitprivacyon: Privacy-preserving mobile gait biometrics using unsupervised learning. *Pattern Recognition Letters*, 161:30–37, 2022.

[39] Vincenzo Dentamaro, Vincenzo Gattulli, Donato Impedovo, and Fabio Manca. Human activity recognition with smartphone-integrated sensors: A survey. *Expert Systems with Applications*, 246:123143, 2024.

[40] Gauri Deshpande, Anton Batliner, and Björn W. Schuller. Ai-based human audio processing for covid-19: A comprehensive overview. *Pattern Recogn.*, 122(C), February 2022.

[41] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[42] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset, 2019.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions, 2020.

[45] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7887–7896, 2020.

[46] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images, 2020.

[47] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Proceedings of the Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[48] Ömer Eskidere, Ahmet Gürhanlı, et al. Voice disorder classification based on multitaper mel frequency cepstral coefficients features. *Computational and mathematical methods in medicine*, 2015, 2015.

[49] Shih-Hau Fang, Yu Tsao, Min-Jing Hsiao, Ji-Ying Chen, Ying-Hui Lai, Feng-Chuan Lin, and Chi-Te Wang. Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, 33(5):634–641, 2019.

[50] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3247–3258. PMLR, 2020.

[51] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video, 2019.

[52] Borko Furht, editor. *Discrete Cosine Transform (DCT)*, pages 203–205. Springer US, Boston, MA, 2006.

[53] Daniel Garcia-Gonzalez, Daniel Rivero, Enrique Fernandez-Blanco, and Miguel R Luaces. A public domain dataset for real-life human activity recognition using smartphone sensors. *Sensors*, 20(8):2200, 2020.

[54] Vincenzo Gattulli, Donato Impedovo, Giuseppe Pirlo, and Lucia Sarcinella. Human activity recognition for the identification of bullying and cyberbullying using smartphone sensors. *Electronics*, 12(2):261, 2023.

[55] General Data Protection Regulation GDPR. General data protection regulation. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*, 2016.

[56] Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du. Advddos: Zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 2023.

[57] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8):128, Jul 2021.

[58] Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8), 2021.

[59] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[61] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[62] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[63] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

[64] Hongzhao Guan and Alexander Lerch. Evaluation of feature learning methods for voice disorder detection. *International Journal of Semantic Computing*, 13(04):453–470, 2019.

[65] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh M. Q. Bui, Marco Fontani, Davide Alessandro Coccomini, Roberto Caldelli, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Giuseppe Amato, Gianpaolo Perelli, Sara Concas, Carlo Cuccu, Giulia Orrù, Gian Luca Marcialis, and Sebastiano Battiato. The face deepfake detection challenge. *Journal of Imaging*, 8(10), 2022.

[66] Rijul Gupta, Dhanshree R Gunjawate, Duy Duong Nguyen, Craig Jin, and Catherine Madill. Voice disorder recognition using machine learning: a scoping review protocol. *BMJ open*, 14(2):e076998, 2024.

[67] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[68] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. *DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame*, pages 255–273. Springer International Publishing, Cham, 2022.

[69] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[70] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[71] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[72] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017.

[73] Molly N Huston, Ira Puka, and Matthew R Naunheim. Prevalence of voice disorders in the united states: a national survey. *The Laryngoscope*, 134(1):347–352, 2024.

[74] Kainat Ibrar, Muhammad Awais Azam, and Muhammad Ehatisham-ul Haq. Personal attributes identification based on gait recognition using smart phone sensors. In *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, pages 94–97, 2020.

[75] Alper Idrisoglu, Ana Luiza Dallora, Peter Anderberg, and Johan Sanmartin Berglund. Applied machine learning techniques to diagnose voice-affecting conditions and disorders: systematic literature review. *Journal of Medical Internet Research*, 25:e46105, 2023.

[76] IEC ISO. Information technology—vocabulary—part 37: Biometrics. *International Standard ISO/IEC*, pages 2382–37, 2022.

[77] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection . In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, Los Alamitos, CA, USA, June 2020. IEEE Computer Society.

[78] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017.

[79] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.

[80] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 852–863. Curran Associates, Inc., 2021.

[81] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[82] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019.

[83] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[84] Samer Y Khamaiseh, Derek Bagagem, Abdullah Al-Alaj, Mathew Mancino, and Hakam W Alomari. Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification. *IEEE Access*, 2022.

[85] Aminollah Khormali and Jiann-Shiun Yuan. Dfdt: an end-to-end deepfake detection framework using vision transformer. *Applied Sciences*, 12(6):2953, 2022.

[86] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recogn.*, 163(C), April 2025.

[87] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12715–12725, 2023.

[88] Aditi Kohli and Abhinav Gupta. Light-weight 3dcnn for deepfakes, faceswap and face2face facial forgery detection. *Multimedia Tools and Applications*, 81(22):31391–31403, 2022.

[89] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[90] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1962–1966, 2018.

[91] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956-1981, Mar 2020.

[92] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[93] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2022.

[94] E.Y. Lam and J.W. Goodman. A mathematical analysis of the dct coefficient distributions for images. *IEEE Transactions on Image Processing*, 9(10):1661–1666, 2000.

[95] E.Y. Lam and J.W. Goodman. A mathematical analysis of the dct coefficient distributions for images. *IEEE Transactions on Image Processing*, 9(10):1661–1666, 2000.

[96] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10117–10127, October 2021.

[97] Congrui Li, Ziqiang Zheng, Yi Bin, Guoqing Wang, Yang Yang, Xuesheng Li, and Heng Tao Shen. Pixel bleach network for detecting face forgery under compression. *IEEE Transactions on Multimedia*, pages 1–13, 2023.

[98] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the 2021 IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition (CVPR), pages 6454–6463, June 2021.

[99] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[100] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2019.

[101] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[102] Xin Liao, Yumei Wang, Tianyi Wang, Juan Hu, and Xiaoshuai Wu. Famm: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

[103] Yu-Jing Lin, Po-Wei Wu, Che-Han Chang, Edward Chang, and Shih-Wei Liao. RelGAN: Multi-Domain Image-to-Image Translation via Relative Attributes . In *Proceedings of the Proceeding of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5913–5921, Los Alamitos, CA, USA, November 2019. IEEE Computer Society.

[104] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[105] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing . In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3677, Los Alamitos, CA, USA, June 2019. IEEE Computer Society.

[106] Songxiang Liu, Haibin Wu, Hung-yi Lee, and Helen Meng. Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 312–319. IEEE, 2019.

[107] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2014.

[108] Jeffrey W Lockhart, Gary M Weiss, Jack C Xue, Shaun T Gallagher, Andrew B Grosner, and Tony T Pulickal. Design considerations for the wisdm smart phone-based sensor mining architecture. In *Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data*, pages 25–33, 2011.

[109] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1683–1698, 2019.

[110] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[111] Luca Maiano, Irene Amerini, Lorenzo Ricciardi Celsi, and Aris Anagnostopoulos. Identification of social-media platform of videos through the use of shared features. *Journal of Imaging*, 7(8):140, 2021.

[112] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints?, 2018.

[113] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019.

[114] Sakorn Mekruksavanich and Anuchit Jitpattanakul. A residual deep learning network for smartwatch-based user identification using activity patterns in daily living. *Computers and Electrical Engineering*, 121:109883, 2025.

[115] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. An overview of privacy-enhancing technologies in biometric recognition. *ACM Comput. Surv.*, 56(12), October 2024.

[116] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3078–3087, 2023.

[117] Pietro Melzi, Hatef Otroshi Shahreza, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Sébastien Marcel, and Christoph Busch. Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 323–331, January 2023.

[118] Pietro Melzi, Ruben Tolosana, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Ivan DeAndres-Tame, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Weisong Zhao, Xiangyu Zhu, Zheyu Yan, Xiao-Yu Zhang, Jinlin Wu, Zhen Lei, Suvidha Tripathi, Mahak Kothari, Md Haider Zama, Debayan Deb, Bernardo Biesseck, Pedro Vidal, Roger Granada, Guilherme Fickel, Gustavo Führ, David Menotti, Alexander Unnervik, Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Parsa Rahimi, Sébastien Marcel, Ioannis Sarridis, Christos Koutlis, Georgia Baltsou, Symeon Papadopoulos, Christos Diou, Nicolò Di Domenico, Guido Borghi, Lorenzo Pellegrini, Enrique Mas-Candela, Ángela Sánchez-Pérez, Andrea Atzori, Fadi Boutros, Naser Damer, Gianni Fenu, and Mirko Marras. Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Information Fusion*, 107:102322, 2024.

[119] Changtao Miao, Qi Chu, Weihai Li, Tao Gong, Wanyi Zhuang, and Nenghai Yu. Towards generalizable and robust face manipulation detection via bag-of-local-feature. *arXiv preprint arXiv:2103.07915*, 2021.

[120] Alakananda Mitra, Saraju P. Mohanty, Peter Corcoran, and Elias Kougianos. A novel machine learning based method for deepfake video detection in social media. In *Proceedings of the 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, pages 91–96, Dec 2020.

[121] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dute, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 2023.

[122] Debadyuti Mukherjee, Riktim Mondal, Pawan Kumar Singh, Ram Sarkar, and Debotosh Bhattacharjee. Ensemconvnet: a deep learning approach for human activity recognition using smartphone sensors for healthcare applications. *Multimedia Tools and Applications*, 79:31663–31690, 2020.

[123] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition, 2017.

[124] Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018.

[125] João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020.

[126] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[127] Pallavi Patil and Kirankumari Patil. A review on disease prediction using image processing. *Journal Electrical and Computer Experiences*, 1(1):18–28, 2023.

[128] Benjamin C Paul, Si Chen, Shaum Sridharan, Yixin Fang, Milan R Amin, and Ryan C Branski. Diagnostic accuracy of history, laryngoscopy, and stroboscopy. *The Laryngoscope*, 123(1):215–219, 2013.

[129] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing, 2016.

[130] Gianpaolo Perelli, Marco Micheletto, Giulia Orru, Giulia Avvisati, Massimo Capozza, and Gian Luca Marcialis. Analysis of human action recognition features in person identification systems for anti-bullying applications. In *Proceedings of the Conference on Human Centred Artificial Intelligence - Education and Practice*, HCAIep '24, page 58, New York, NY, USA, 2024. Association for Computing Machinery.

[131] Gianpaolo Perelli, Andrea Panzino, Roberto Casula, Marco Micheletto, Giulia Orrù, and Gian Luca Marcialis. Vulnerabilities in machine learning-based voice disorder detection systems. In *Proceedings of the 2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2024.

[132] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics (SIGGRAPH'03)*, 22(3):313–318, 2003.

[133] Hai X. Pham, Yuting Wang, and Vladimir Pavlovic. Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network, 2018.

[134] Stanislav Pidhorskyi, Donald Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders, 2020.

[135] Maria E Powell, Marcelino Rodriguez Cancio, David Young, William Nock, Beshoy Abdelmessih, Amy Zeller, Irvin Perez Morales, Peng Zhang, C Gaelyn Garrett, Douglas Schmidt, et al. Decoding phonation with artificial intelligence (dep ai): proof of concept. *Laryngoscope Investigative Otolaryngology*, 4(3):328–334, 2019.

[136] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 86–103, Cham, 2020. Springer International Publishing.

[137] Yao Qin, Nicholas Carlini, Ian J. Goodfellow, G. Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. *ArXiv*, abs/1903.10346, 2019.

[138] Lawrence Rabiner and Ronald Schafer. *Theory and Applications of Digital Speech Processing*. Prentice Hall Press, USA, 1st edition, 2010.

[139] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.

[140] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos, F. De Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. Venkatesh, J. M. Singh, G. Li, L. Bergeron, S. Isadskiy, R. Ramachandra, C. Rathgeb, D. Frings, U. Seidel, F. Knopjes, R. Veldhuis, D. Maltoni, and C. Busch. Morphing attack detection - database, evaluation platform and benchmarking. *IEEE Trans. on Information Forensics and Security*, 2020.

[141] Elangovan Ramanujam, Thinagaran Perumal, and S Padmavathi. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 21(12):13029–13040, 2021.

[142] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[143] Christian Rathgeb, Naser Damer, Kiran Raja, Ulrich Scherhag, and Christoph Busch, editors. *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks.* Springer International Publishing, Cham, 2022.

[144] Jonathan Reid, Preet Parmar, Tyler Lund, Daniel K Aalto, and Caroline C Jeffery. Development of a machine-learning based voice disorder screening tool. *American Journal of Otolaryngology*, 43(2):103327, 2022.

[145] Sergio Romero-Tapiador, Blanca Lacruz-Pleguezuelos, Ruben Tolosana, Gala Freixer, Roberto Daza, Cristina M Fernández-Díaz, Elena Aguilar-Aguilar, Jorge Fernández-Cabezas, Silvia Cruz-Gil, Susana Molina, Maria Carmen Crespo, Teresa Laguna, Laura Judith Marcos-Zambrano, Ruben Vera-Rodriguez, Julian Fierrez, Ana Ramírez de Molina, Javier Ortega-Garcia, Isabel Espinosa-Salinas, Aythami Morales, and Enrique Carrillo de Santa Pau. Ai4fooddb: a database for personalized e-health nutrition and lifestyle through wearable devices and artificial intelligence. *Database*, 2023:baad049, 07 2023.

[146] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[147] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

[148] Salenna Russello. The impact of media exposure on self-esteem and body satisfaction in men and women. *Journal of Interdisciplinary Undergraduate Research*, 1(1):Article 4, 2009.

[149] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[150] Nazir Saleheen, Md Azim Ullah, Supriyo Chakraborty, Deniz S Ones, Mani Srivastava, and Santosh Kumar. Wristprint: Characterizing user re-identification risks from wrist-worn accelerometry data. In *Proceedings of the*

2021 ACM SIGSAC Conference on Computer and Communications Security, pages 2807–2823, 2021.

[151] Alejandro Sanchez Guinea, Simon Heinrich, and Max Mühlhäuser. Activity-free user identification using wearables based on vision techniques. *Sensors*, 22(19):7368, 2022.

[152] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019.

[153] Hatef Otroshi Shahreza, Christophe Ecabert, Anjith George, Alexander Unnervik, Sébastien Marcel, Nicolò Di Domenico, Guido Borghi, Davide Maltoni, Fadi Boutros, Julia Vogel, Naser Damer, Ángela Sánchez-Pérez, Enrique Mas-Candela, Jorge Calvo-Zaragoza, Bernardo Biesseck, Pedro Vidal, Roger Granada, David Menotti, Ivan DeAndres-Tame, Simone Maurizio La Cava, Sara Concas, Pietro Melzi, Ruben Tolosana, Ruben Vera-Rodriguez, Gianpaolo Perelli, Giulia Orrù, Gian Luca Marcialis, and Julian Fierrez. Sdfr: Synthetic data for face recognition competition. In *Proceedings of the 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2024.

[154] Hatef Otroshi Shahreza and Sébastien Marcel. Hyperface: Generating synthetic face recognition datasets by exploring face embedding hypersphere. In *Proceedings of the The Thirteenth International Conference on Learning Representations*, 2025.

[155] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing, 2020.

[156] Taoran Sheng and Manfred Huber. Weakly supervised multi-task representation learning for human activity analysis using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–18, 2020.

[157] Thomas Sikora. The mpeg-4 video standard verification model. *IEEE Transactions on circuits and systems for video technology*, 7(1):19–31, 1997.

[158] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022.

[159] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 919–925. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[160] Nawel Souissi and Adnane Cherif. Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In *Proceedings of the 2015 7th international conference on modelling, identification and control (ICMIC)*, pages 1–6. IEEE, 2015.

[161] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[162] Marcin Straczkiewicz, Peter James, and Jukka-Pekka Onnela. A systematic review of smartphone-based human activity recognition methods for health research. *NPJ Digital Medicine*, 4(1):148, 2021.

[163] Giuseppe Stragapede, Ruben Vera-Rodriguez, Ruben Tolosana, Aythami Morales, Alejandro Acien, and Gaël Le Lan. Mobile behavioral biometrics for passive authentication. *Pattern Recognition Letters*, 157:35–41, 2022.

[164] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *Proceedings of the European conference on computer vision*, pages 111–127. Springer, 2022.

[165] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4), jul 2017.

[166] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[167] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[168] Adane Nega Tarekegn, Mohib Ullah, and Faouzi Alaya Cheikh. Deep learning for multi-label learning: A comprehensive survey. *arXiv preprint arXiv:2401.16549*, 2024.

[169] João Paulo Teixeira, Paula Odete Fernandes, and Nuno Alves. Vocal acoustic analysis–classification of dysphonic voices with artificial neural networks. *Procedia computer science*, 121:19–26, 2017.

[170] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3735–3744, 2017.

[171] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Proceedings of the Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI*, page 716–731, Berlin, Heidelberg, 2020. Springer-Verlag.

[172] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[173] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 442–456, Cham, 2021. Springer International Publishing.

[174] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[175] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[176] Hirohito Umeno, Masamitsu Hyodo, Tomoyuki Haji, Hirotaka Hara, Mitsuyoshi Imaizumi, Miyoko Ishige, Masanobu Kumada, Kiyoshi Makiyama, Noriko Nishizawa, Koichiro Saito, et al. A summary of the clinical practice guideline for the diagnosis and management of voice disorders, 2018 in japan. *Auris Nasus Larynx*, 47(1):7–17, 2020.

[177] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society*, 2(3):128–145, 2021.

[178] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[179] G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[180] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[181] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003.

[182] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17:151–178, 2020.

[183] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019. International Joint Conferences on Artificial Intelligence Organization, August 2019.

[184] Liying Yang. Classifiers selection for ensemble learning based on accuracy and diversity. *Procedia Engineering*, 15:4266–4270, 2011.

[185] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He. Masked relation learning for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1696–1708, 2023.

[186] Michael Yeung, Toya Teramoto, Songtao Wu, Tatsuo Fujiwara, Kenji Suzuki, and Tamaki Kojima. Variface: Fair and diverse synthetic dataset generation for face recognition, 2025.

[187] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.

[188] Xi Yin and Xiaoming Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, 2017.

[189] Nada R Yousif, Hossam Magdy Balaha, Amira Y Haikal, and Eman M El-Gendy. A generic optimization and learning framework for parkinson disease via speech and handwritten records. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):10673–10693, 2023.

[190] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[191] Haoyu Zhang, Sushma Krupa Venkatesh, Raghavendra Ramachandra, K. Bommanna Raja, Naser Damer, and Christoph Busch. Mip-gan—generating strong and high quality morphing attacks using identity prior driven gan. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3:365–383, 2021.

[192] Jian Zhang, Jiangqun Ni, and Hao Xie. Deepfake videos detection using self-supervised decoupling network. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2021.

[193] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2578–2593, 2020.

[194] Tao Zhang. Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5):6259–6276, 2022.

[195] Yihua Zhang, Jinyang Qian, Xiaojun Zhang, Yishen Xu, and Zhi Tao. Pathological voice detection using joint subsapce transfer learning. *Applied Sciences*, 12(16), 2022.

[196] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

[197] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[198] Yuan Zhao, Bo Liu, Ming Ding, Baoping Liu, Tianqing Zhu, and Xin Yu. Proactive deepfake defence via identity watermarking. In *Proceedings of the*

2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 4591–4600, Jan 2023.

[199] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 86–107, 2021.

[200] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9299–9306, Jul. 2019.

[201] Xin Zhou, Yongtao Wang, and Peihan Wu. Detecting deepfake videos via frame serialization learning. In *Proceedings of the 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, pages 391–395, Nov 2020.

[202] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[203] Ángel Fernández Gambín, Anis Yazidi, Athanasios Vasilakos, Hårek Haugerud, and Youcef Djenouri. Deepfakes: current and future trends. *Artificial Intelligence Review*, 57(3), 2024.

Acknowledgements

This doctoral thesis was conducted within the PhD Programme in Electronic and Computer Engineering at the University of Cagliari (Cycle XXXVIII), supported by a scholarship co-funded by the Italian Ministry of University and Research (MUR) – Decree No. 352 of 9.4.2022, under the PNRR - funded by the European Union – NextGenerationEU – Mission 4 "Education and Research", Component 2 "From Research to Business", Investment 3.3, and by the company Wallife S.r.l..

The journey of a PhD is never a solitary endeavour. It is a path walked with the guidance, support, and encouragement of many people, to whom I wish to express my deepest gratitude.

First and foremost, I extend my sincere thanks to my supervisor, Prof. Gian Luca Marcialis, for his guidance, expertise, patience, and constancy. I am equally grateful to my co-supervisor, Prof. Giovanni Puglisi, for his invaluable insights and constant support.

My heartfelt thanks go to all the members of the PRA Lab research group. I wish to thank the whole group for the stimulating discussions and the collaborative spirit, and for making the lab a joyful and productive environment.

Finally, my deepest gratitude goes to my family. To my partner, who always pushes me to do my best and who rolled up her sleeves during my time abroad. To our beloved daughter, who fills even the most difficult days with joy. To my parents, for their unconditional love and unwavering faith in me.